

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/104723>

**Copyright and reuse:**

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

# THE BRITISH LIBRARY

BRITISH THESIS SERVICE

**TITLE** ADAPTIVE-RATE DIGITAL SPEECH  
TRANSMISSION.

**AUTHOR** Wei  
HE

**DEGREE** Ph.D

**AWARDING  
BODY** Warwick University

**DATE** 1993

**THESIS  
NUMBER** DX182709

THIS THESIS HAS BEEN MICROFILMED EXACTLY AS RECEIVED

The quality of this reproduction is dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction. Some pages may have indistinct print, especially if the original papers were poorly produced or if awarding body sent an inferior copy. If pages are missing, please contact the awarding body which granted the degree.

Previously copyrighted materials (journals articles, published texts etc.) are not filmed.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no information derived from it may be published without the author's prior written consent.

Reproduction of this thesis, other than as permitted under the United Kingdom Copyright Designs and Patents Act 1988, or under specific agreement with the copyright holder, is prohibited.

# Adaptive-Rate Digital Speech Transmission

By  
*Wei He. BSc.*

Submitted for the Degree of Doctor of Philosophy  
to the Higher Degrees Committee  
University of Warwick

Department of Engineering  
University of Warwick.

April 1993

*To*  
*My Father and Mother;*  
*My Wife Lan and Son Kaowen.*

谨以此献给我的父亲母亲，我的妻子和儿子。

# Contents

<b>Contents</b>	<b>i</b>
<b>List of figures</b>	<b>iv</b>
<b>List of tables</b>	<b>viii</b>
<b>Acknowledgment</b>	<b>x</b>
<b>Declaration</b>	<b>xi</b>
<b>Abstract</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Speech Sources . . . . .	2
1.2 Speech Quality . . . . .	6
1.3 Statement of Problems . . . . .	8
1.4 Organisation of Thesis . . . . .	13
<b>2 Review of Speech Compression Techniques</b>	<b>16</b>
2.1 Introduction . . . . .	17
2.2 Analog Speech Bandwidth Reduction . . . . .	19
2.2.1 Frequency Division . . . . .	19
2.2.2 Time Domain Speech Gapping . . . . .	20
2.3 Waveform Encoding Techniques . . . . .	23
2.3.1 Pulse Code Modulation (PCM) . . . . .	23
2.3.2 Differential Pulse Code Modulation (DPCM) . . . . .	25
2.3.3 Delta Modulation (DM) . . . . .	26
2.3.4 Sub-Band Encoding (SBC) . . . . .	27
2.3.5 Transform Coding (TC) . . . . .	29

2.4	Parameter Encoding Technique . . . . .	30
2.4.1	Voiced/Unvoiced/Silence and Pitch Detection . . . . .	32
2.4.2	Linear Predictive Coding (LPC) . . . . .	36
2.4.3	Channel Vocoders . . . . .	38
2.4.4	Formant Vocoders . . . . .	40
2.5	Mid-Band Encoding Techniques . . . . .	42
2.5.1	Combine Waveform Encoding with Data Compression . . . . .	43
2.5.2	Advanced Waveform Encoding . . . . .	44
2.5.3	Simplified Parameter Encoding Schemes . . . . .	45
2.6	Comparison of Speech Coding Techniques . . . . .	46
<b>3</b>	<b>Adaptive-Rate Sampling for Speech Compression</b>	<b>48</b>
3.1	Introduction . . . . .	49
3.2	Shannon Sampling Theorem . . . . .	51
3.3	Optimal Sampling Rate Algorithm for Multi-Band-Pass Signals	55
3.3.1	Time Domain Compression Technique . . . . .	55
3.3.2	Frequency Companding Technique . . . . .	60
3.3.3	Comparison of Time Domain Compression and Fre- quency Companding Techniques . . . . .	62
3.4	Extraction of a Multi-Band-Pass Signal from a Speech Signal . . . . .	64
3.4.1	Threshold Selection . . . . .	66
3.4.2	Relation Between Eliminated Spectrum and Speech Quality . . . . .	70
3.5	Adaptive-Rate Sampling System Design . . . . .	73
3.5.1	System Description . . . . .	73
3.5.2	Simulation Results . . . . .	76
<b>4</b>	<b>Reconstruction of Speech from Its Frame Differences</b>	<b>83</b>
4.1	Introduction . . . . .	84
4.2	Examination of Time-Variation of Speech Signal . . . . .	85
4.3	Compression Technique . . . . .	88

4.4	System Description . . . . .	94
4.4.1	Voiced/Unvoiced/Silence and Pitch Detector . . . . .	96
4.4.2	Compression Processor . . . . .	101
4.4.3	Reconstruction Processor . . . . .	104
4.5	Simulation Results . . . . .	107
<b>5</b>	<b>Speech Scrambling Employing Adaptive-Rate Sampling</b>	<b>113</b>
5.1	Introduction . . . . .	114
5.2	A Brief Review of Speech Scrambling Techniques . . . . .	117
5.2.1	One Dimensional Speech Scrambling . . . . .	117
5.2.2	Two Dimensional Speech Scramblers . . . . .	123
5.3	Speech Scrambling Using Adaptive-Rate Sampling Algorithm .	124
5.4	Scrambling System Description . . . . .	130
5.4.1	Single User System . . . . .	131
5.4.2	Multi-User System . . . . .	132
5.4.3	Sampling Rate Tolerance . . . . .	137
5.5	Subjective Tests to Measure Residual Intelligibility . . . . .	139
5.6	Simulation Results . . . . .	142
<b>6</b>	<b>Experimental Results and Practical Implementations</b>	<b>145</b>
6.1	Introduction . . . . .	146
6.2	Combine ARS with Speech Coding . . . . .	147
6.2.1	Comparison of Combined Technique with Low Bit Rate PCM and Low Rate Sampling . . . . .	148
6.2.2	Combine Low Rate Sampling with Hadamard Coding .	154
6.2.3	Combine Low-Rate Sampling with Adaptive PCM . . .	159
6.3	Combine ARS and Channel Coding for Reliable Transmission	163
6.3.1	System Description . . . . .	165
6.3.2	Simulation Result . . . . .	168
6.4	Application of ARS in Speech Storage . . . . .	170
6.5	Real-Time ARS Using DSP . . . . .	175

---

<b>7 Conclusion and Suggestions for Further Work</b>	<b>179</b>
7.1 Conclusion . . . . .	180
7.1.1 Adaptive-Rate Sampling for Speech Transmission . . .	180
7.1.2 Low Rate Speech Transmission Using Frame Differences	181
7.1.3 Speech Scrambling Using Adaptive-Rate Sampling . . .	182
7.2 Suggestions for Further Studies . . . . .	183
7.2.1 Combine Low Rate Sampling With Parameter Coding .	183
7.2.2 Speech Compression Using Differential Signal in Time Domain . . . . .	184
7.2.3 Reliable Transmission of Speech . . . . .	185
7.2.4 Noise Reduction Using Dynamic Thresholding . . . . .	186
 Bibliography	 188
 List of Symbols and Abbreviations	 202
 Appendix	 205



## List of Figures

1.1	The principal of human vocal organs ([Parsons, 87]) . . . . .	3
1.2	Acoustic waveform for the phrase of 'HF systems are' . . . . .	5
1.3	3D spectrungraph for the phrase of 'HF systems are' . . . . .	5
1.4	Speech production model . . . . .	6
2.1	Speech waveform for the phrase 'five members'. . . . .	22
2.2	Simplified block diagram of a PCM system . . . . .	24
2.3	Simplified block diagram of a DM system . . . . .	26
2.4	Block diagram of sub-band encoding system . . . . .	28
2.5	Excitation model of speech generation . . . . .	31
2.6	Block diagram of the basic elements of a parameter coding system . . . . .	31
2.7	Speech production model. . . . .	37
2.8	Simplified block diagram of a channel vocoder. . . . .	39
2.9	Simplified block diagram of a formant vocoder . . . . .	41
2.10	Waveform for adaptive 2-bit PCM system . . . . .	43
2.11	Simplified block diagram of APC transmitter . . . . .	45
2.12	Bit rates of speech encoders . . . . .	47
3.1	Spectrum of sampled low pass signal . . . . .	52
3.2	Spectrum of band-pass function . . . . .	54
3.3	Example of sub-Nyquist sampling . . . . .	59
3.4	Frequency Companding of the speech signal . . . . .	61
3.5	Elimination of redundancies from speech spectrum . . . . .	65

3.6	Waveform of the word of 'city' . . . . .	68
3.7	Spectrum for voiced and unvoiced sound in the word of 'city' .	68
3.8	Relation between threshold level and retained bandwidth . . .	72
3.9	Relation between retained bandwidth and intelligibility . . . .	72
3.10	System block diagram for Time Domain Compression technique	74
3.11	System block diagram for Frequency Companding technique .	74
3.12	Simulation result for Time Domain Compression system . . .	79
3.13	Simulation result for Frequency Companding system . . . . .	80
4.1	Time waveform of the phrase of 'HF systems' . . . . .	87
4.2	A short time segment of speech signal (32 ms) . . . . .	90
4.3	Speech waveform and frequency responses in a fixed length frame (32 ms) . . . . .	91
4.4	Differential signal in fixed length frame . . . . .	92
4.5	Speech waveform and frequency response in two consecutive periodic-frames . . . . .	93
4.6	Differential signal and reconstructed signal . . . . .	94
4.7	System block diagram . . . . .	95
4.8	Block diagram of SIFT algorithm . . . . .	98
4.9	Typical signals from the SIFT algorithm ([Markel:72]) . . . .	100
4.10	Flow chart of compression process . . . . .	102
4.11	Flow chart of reconstruction process . . . . .	105
4.12	Reconstructed unvoiced speech . . . . .	108
4.13	Reconstructed voiced speech . . . . .	110
4.14	Simulation result of reconstruction of speech from frame dif- ferences . . . . .	112
5.1	Frequency inversion . . . . .	118
5.2	Bandsplitting . . . . .	119
5.3	Bandsplitting combined with frequency inversion . . . . .	120
5.4	Frequency inversion followed by cyclic bandshift . . . . .	120

5.5	Time inversion . . . . .	121
5.6	Time segment permutation (TSP) . . . . .	122
5.7	Time frequency segment permutation (TFSP) . . . . .	124
5.8	Spectral shifting and inversion . . . . .	126
5.9	Spectral re-ordering and translation . . . . .	126
5.10	Spectral re-ordering and translation at different sampling rate	127
5.11	Example of the scrambling procedure . . . . .	129
5.12	Block diagram of scrambling system . . . . .	131
5.13	Scrambled spectrum arrangement (with system bandwidth of 20 kHz) . . . . .	133
5.14	Multi user speech scrambling system . . . . .	134
5.15	Communications resource plane in three users system . . . . .	136
5.16	Descrambling sampling frequency tolerance for various system bandwidth . . . . .	139
5.17	Three scrambling patterns for 8 sub-bands . . . . .	140
5.18	Three scrambling patterns for 16 sub-bands . . . . .	140
5.19	A simulation result of 4 sub-band scrambling technique . . . . .	143
5.20	A detailed illustration of the result of the scrambling technique	144
6.1	Intelligibility test for low rate sampling and low bit rate PCM	150
6.2	Segmental signal-to-noise ratio for low rate sampling and low bit rate PCM . . . . .	151
6.3	Spectral distortion for low rate sampling and low bit rate PCM	151
6.4	Intelligibility test for combined system . . . . .	153
6.5	Segmental signal-to-noise ratio for combined system . . . . .	153
6.6	Spectral distortion for combined system . . . . .	154
6.7	Intelligibility test for the combined low-rate sampling and APCM system . . . . .	162
6.8	Segmental signal-to-noise ratio for the combined low-rate sam- pling and APCM system . . . . .	162

---

6.9	Spectrum distortion for the combined low-rate sampling and APCM system . . . . .	163
6.10	Simplified block diagram of combined adaptive-rate sampling and error control coding technique . . . . .	166
6.11	1/2 rate convolutional coder ( $k=3$ ) . . . . .	167
6.12	The deleting map for the rate of 3/4 convolutional code . . . .	168
6.13	The performance of the three transmission formats as a func- tion of channel SNR . . . . .	169
6.14	Speech waveform in two 16-ms frames. . . . .	173
6.15	Compressed speech waveform in two 16-ms frames. . . . .	174
6.16	DSP timing diagram . . . . .	178

## List of Tables

3.1	Speech status and the optimal sampling rate for each frame . .	82
4.1	Test results of SIFT algorithm . . . . .	101
6.1	The relationship of sampling rate $w$ , $N$ , $L$ and $l$ ( $m = 8$ ) . . . .	158
6.2	Step size multipliers for $B = 2, 3$ and 4. . . . .	161
6.3	Sampling rate, source and channel code formats. . . . .	165
6.4	The samples of the speech signal for frame 1 and 2 . . . . .	173
6.5	The samples of the compressed speech signal for frame 1 and 2.	174
6.6	The samples of the compressed speech signal for frame 1 and 2 in Hex format. . . . .	175

## Acknowledgment

I would like to take this opportunity to express my sincere gratitude to my supervisor, Prof. B. Honary, for his guidance, constant encouragement, and valuable comments through the course of this research. Thanks are due to Prof. M. Darnell of Hull University, Dr. M. Dodson and Dr. M. Beaty of York University for their valuable advice and suggestions.

I also like to express my appreciation to my colleagues in Hull-Lancaster (former Hull-Warwick) Communications Research Group for their friendship and assistances.

Finally I must thank my parents, my wife Lan and my son Kaowen for their love, support and constant encouragement,

## Declaration

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy by the Higher Degree Committee at the University of Warwick. The thesis has been composed and written based on the research undertaken by the author. The research materials have not been submitted in any previous application for a higher degree. All sources of information are specifically acknowledged in the content.

## Abstract

This thesis investigates adaptive-rate sampling (ARS) algorithms for speech communications. The sampling algorithms allow a multi-band-pass signal to be sampled at a rate much lower than the Nyquist rate without causing significant speech degradation. These ARS techniques demonstrate a capability of efficient speech transmission and storage, and provide useful methods for reliable speech transmission over time-variable channels. Experimental results show that without using any speech coding technique, the ARS produces speech at a transmission rate of 20 kb/s with a quality similar to that produced by 6-bit PCM which requires a transmission rate of 48 kb/s. The ARS techniques are easy to be combined with low bit rate speech coding techniques to achieve further speech compression.

A speech signal manipulation technique for speech compression is also investigated in this thesis. This technique, namely reconstruction of speech from its frame differences, extracts the differences from two consecutive frames of speech and transmits them. At the receiver, the differences are used to modify the previously reconstructed frame signal to produce the current frame speech. Due to the quasi-periodic property in speech signal, the differential signal usually occupies a very narrow bandwidth. By applying the ARS to the differential signal, a significant sampling rate reduction is achieved.

Secure speech transmission can be established by employing the ARS algorithm. A new speech scrambling technique is presented in this thesis. The speech is spectrally scrambled at the transmitter to produce a clearly defined multiple band pass structure. The reconstruction 'key' comprises a knowledge of the spectral frequency bands of the scrambled elements together with a unique sampling rate which will allow correct recombination of these elements. In practice, the bandpass structure and sampling rate would be changed regularly according to an appropriate key sequence to prevent unauthorised reconstruction.



# **Chapter 1**

## **Introduction**

## 1.1 Speech Sources

Speech is human's most common form of communication. Speech sounds are produced by air from the lungs exciting the vocal tract, a time varying nonuniform cavity extending from the vocal cords to the lips, with occasional coupling of the nasal cavity by means of velum [Rabiner,*et al.*(1978)]. Speech has several distinct types of acoustic waveforms: quasi random, quasi periodic, quiescent period and bursts of energy. This formation is generated by the anatomical coordination of the tongue, teeth, jaw, nasal cavity, velum, vocal cords, and air flow generated by the lung cavity [Parsons,(1987)]. This anatomical structure is shown in Figure 1.1. Speech sound can be classified into three distinct classes according to their mode of excitation:

- (i) **Voiced** sound, which normally include all vowels and some consonants, such as [m,n,l,w] [Holmes,(1988)] are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulse of air which excite the vocal tract. The fundamental frequency of this signal lies typically in the range 50 - 200 Hz for adult male speakers, and about 160 - 400 Hz for adult females [Rabiner,*et al.*(1978)].
- (ii) **Fricative** sound are generated by forming a constriction at some point in the vocal tract (usually toward the mouth end), and forcing air through the constriction at a high enough velocity to pro-

duce turbulence. This creates a broad spectrum noise source to excite the vocal tract.

- (iii) **Plosive** sounds result from making a complete closure (again, usually toward the front of the vocal tract), building up pressure behind the closure, and abruptly releasing it.

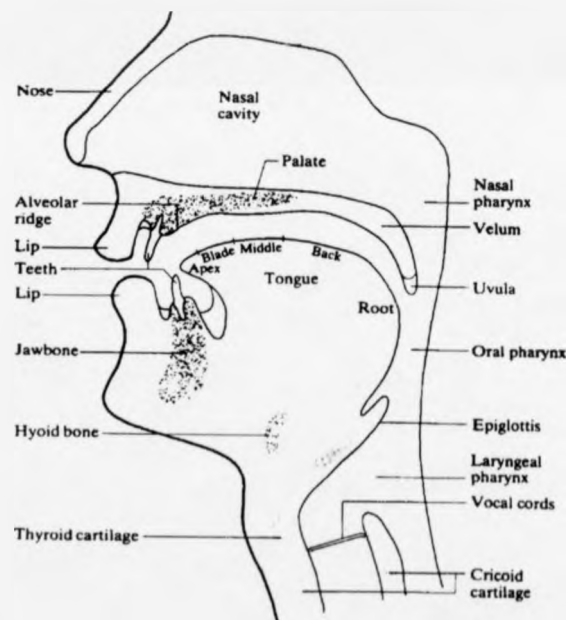


Figure 1.1: The principal of human vocal organs ([Parsons, 87])

In speech signal processing, these types of speech are classified as voiced speech, unvoiced speech (fricative and plosive sound) and silent (no speech).

These classes of speech are shown in Figure 1.2 which is the speech waveform for the partial sentence of 'HF systems are for users to use'. It shows that the speech properties change during the transition from one class to another. For example, if the excitation changes between voiced and unvoiced speech, there are larger changes in peak amplitude of the signal and there is considerable variation of fundamental frequency. The same piece of speech is displayed in three dimensional spectrumgraph in Figure 1.3. The spectrum changes dramatically from voiced speech to unvoiced speech. During a voiced section, the speech energy is concentrated in certain frequency ranges and forms several peaks in its spectrum. These peaks, corresponding to resonances in the vocal tract, are called formants in speech processing [Rabiner,*et al.*(1978)]. The number and location of the formants vary from sound to sound, and are different for different speakers. The spectrumgraph exhibits the formants pattern during voiced speech and broad band noise-like spectrum during unvoiced speech.

A model of a speech production process which has received wide acceptance is illustrated in Figure 1.4. Voiced excitation is modelled by a train of unipolar, unit amplitude impulses at the desired pitch frequency. Unvoiced excitation is modelled as the output from a pseudo-random noise generator; the voiced/unvoiced switch selects the appropriate excitation.

Although the speech waveform varies from time to time, it changes little for a short time segment [Rabiner,*et al.*(1978)]. Most speech processing techniques isolate such segments and process them as if they were short

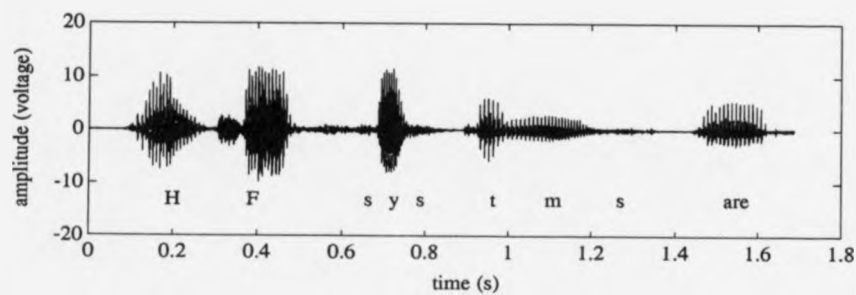


Figure 1.2: Acoustic waveform for the phrase of 'HF systems are'

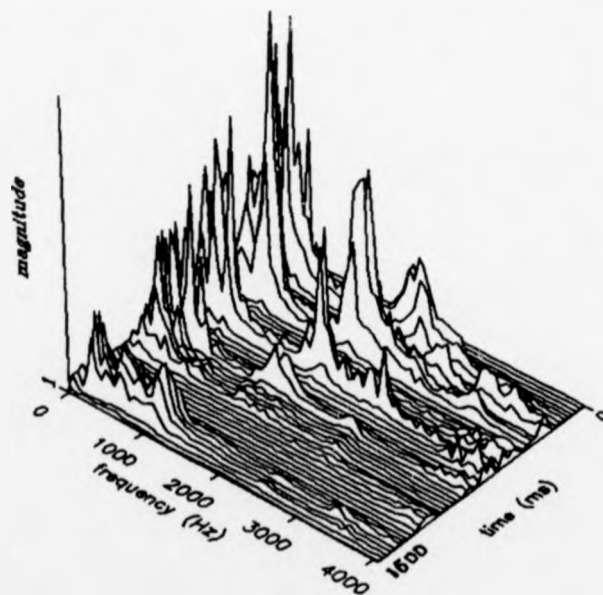


Figure 1.3: 3D spectrumgraph for the phrase of 'HF systems are'

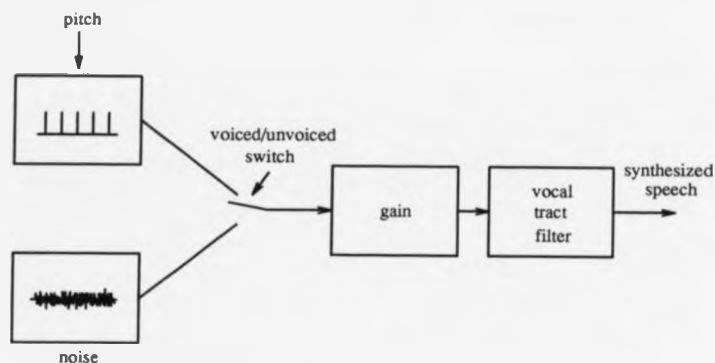


Figure 1.4: Speech production model

segments of sustained sound having fixed properties [Rabiner,*et al.*(1978)]. These segments, sometime called frames, are typically over the time intervals of 10 to 30 ms [Rabiner,*et al.*(1978)] and have the appearance either of a low-energy random (unvoiced) signal, such as /*tf*/ in 'H', or a high-energy pseudo-periodic (voiced), such as /*a:*/ in 'are' (Figure 1.2).

## 1.2 Speech Quality

In communication systems, speech quality is determined by many attributes. The most important of these are listed as follows [Gieseler,*et al.*(1980)]:

- (i) **Intelligibility.** Intelligibility indicates the ability of a system to correctly transmit understandable speech (words and sentences). Most intelligibility measures are based on the responses of a jury of listeners. Typically, a list of understandable words and sentences

is transmitted through the channel under test and the output is played back for the jury; the intelligibility measure is based on the number of correct identifications by the jury. The list of words and sentences must contain all the phonemes of the language in proportion close to their natural frequency of occurrence in speech [Parsons,(1987)]. Although intelligibility is a necessary condition for a speech system, it is not a sufficient condition for all speech communication systems [Kitawaki.(1988)].

- (ii) **Speaker recognition.** The ability of a speech system to allow a listener to recognise the speaker he/she knows is another important attribute. This is an important condition for some security systems and military communications.
- (iii) **Naturalness.** This is closely related to speaker recognition. It is the attribute of sounding natural - like a human utterance, not machine-made. A natural sound is important in public telephone systems.
- (iv) **Lack of noise and distortion.** Background noise and distortion of the speech signal can degrade the speech quality. Noise is usually perceived as a constant hissing sound while distortion may produce a rasp-like sound or echoes during speech but disappears when the speech is not present.

- (v) **Echoes.** Echoes are the presence of audible sounds which echo the original sound. They are a form of distortion usually caused by impedance mismatching in telephone systems. Echoes can also give rise to a hollow sound in speech often described as 'speaking in a barrel'.
- (vi) **Interference.** This degradation is caused by unwanted signals getting into the wrong frequency band or channel. The sources of interference are usually man-made. It may be perceived as a tone, noise, buzz or clicks [Parsons,(1987)].

### 1.3 Statement of Problems

Digital speech transmission has become significantly popular in the modern communications systems. There are many advantages that can be gained by representing speech signals in digital form.

- (i) The degradation in long-distance transmission can be significantly reduced by using regenerative repeaters.
- (ii) The modern programmable digital signal processor has made it very easy to handle the speech signals in digital form.
- (iii) If a speech signal is in digital form it is identical to data of other form. Thus the speech and data can be integrated into one transmission network with no need to distinguish between them except



in the decoding.

- (iv) In some communication systems, security is required to prevent any unauthorized reconstruction of speech, such as in military communications. The digital representation has a distinct advantage over an analog system. For secrecy, the information bits can be scrambled in a manner which can ultimately be unscrambled at the receiver.
- (v) Digital systems are reliable and very compact. Very Large Scale Integrated circuit technology has advanced to a state where extremely complex systems can be implemented on a single chip. Logic speeds are fast enough so that the tremendous number of computations required in many signal processing functions can be implemented in real-time at speech sampling rates. For these and other reasons digital techniques are being increasingly applied in the speech communications problem [Flanagan,(1976)].

To make use of digital equipment in speech processing, the analogue speech must be digitised before any further processing. This is done by means of sampling and encoding. If speech is digitised by means of conventional analog-to-digital conversion (A/D) techniques, a very large number of bits must be transmitted or stored. The total number of bits depends upon both sampling rate and the number of bits used to represent the amplitude of each sample. For conventional sampling theory, in order to reconstruct a

signal without loss of information, the sampling rate must satisfy the Nyquist criterion, which says the minimum sampling rate must not be less than twice the maximum frequency of the signal. In telecommunication systems, the typical bandwidth for speech signals is 4000 Hz, so the Nyquist rate is 8000 Hz. If 8 bits are used to represent the amplitude of each sample, the bit rate is 64,000 bits per second. This bit rate is usually prohibitive. Speech compression refers to reducing this bit rate to achieve an economical processing.

Ideally, we would like a compression system to provide a very good quality of speech at very low bit rate with reliable low-cost equipment. However, there is at present no way of satisfying the user on all of these points, and there must inevitably be a compromise among these conflicting attributes of speech quality, bit rate and equipment complexity. The relative importance of these attributes depends on the application.

Speech compression has been achieved by means of low rate speech coding techniques. Speech coding techniques can be grouped into three classes, waveform, transform and parametric coding. Waveform coding interprets speech as a bit stream that can be approximately reconstructed into the original waveform for a variety of signals. It may use the waveform's statistical properties, such as bandwidth, amplitude distribution, autocorrelation, and power spectral density. Transform coding takes mathematical transformation, such as Fourier transform, Hadamard transform, to blocks of speech samples. The transform coefficients are then encoded and transmitted. Parametric coding extracts certain parameters from the original speech, and en-

codes them as digital bits. When speech is reconstructed, the waveform may not appear to be similar as the original, but it will sound similar.

The bandwidth or bit rate needed to transmit a speech signal directly affects the speech quality. High speech quality generally requires high bandwidth or bit rate. Speech quality is divided into four categories [Jayant, *et al.*(1990)]

- (i) Synthetics quality has bit rates below 4.8 kb/s and sounds rather mechanical.
- (ii) Communication quality has rates ranging from 4.8 to 16 kb/s, produces speech that is intelligible, but noticeably distorted.
- (iii) Toll speech quality has rates between 16 and 64 kb/s and results in speech that is comparable to that of an analog telephone network; it has a bandwidth of 4000 Hz.
- (iv) Commentary quality is reached at rates exceeding 64 kb/s. Although it sounds similar to toll quality speech, it has a large bandwidth.

Waveform coding methods usually produce speech that tends to be at least toll quality at the price of high bit rate (16 kb/s above), Transform coding can provide a speech transmission with communication quality at about 9.6 kb/s [Holmes,(1988)]. Parametric coding method produce speech of poorer quality but at lower bit rate. Parametric coding is the initial choice of many applications because of the low bit rates.

The conventional speech encoding techniques attempt to make more efficient use of the communication channel by reducing the inherent redundancy in the speech source to give a lower transmission rate. However, less attention has been given to the possibility of improving the standard sampling rate. As mentioned earlier, the conventional approach to determining the sampling rate for a given analogue signal has been based on the Nyquist criterion: this relates the sampling rate to the maximum frequency, or to the difference of the maximum and minimum frequencies of the analogue signal to be sampled. Because the transmission rate depends upon both sampling rate and the number of bits used to represent the amplitude of each sample, if a lower sampling rate can be used to sample the analog signal, less samples are needed to be processed. Combined with the existing low bit rate speech coding techniques, the overall transmission rate can be reduced significantly.

In this thesis, the author first develops the new techniques which minimize the sampling rate for generalised multi-band-pass signals. These techniques are then applied to speech signals to achieve an optimal sampling rate. Here, speech signal is analysed in the frequency domain. A threshold is employed to eliminate some of the trivial frequency components, a multi-band-pass signal can be obtained and the optimum sampling algorithms which were developed at first stage can be employed to achieve low sampling rate processing. Because of the speech property of formants, the total loss of energy in the processing of filtering out those trivial frequency components will be little, therefore the quality of speech will not be affected significantly. The

quasi-periodical property of voiced speech gives a possibility of further sampling rate reduction. A speech signal manipulation technique is developed. This technique takes the advantage of the fact that two consecutive frames of voiced speech are highly correlated. It is possible to reconstruct one frame signal by slightly modifying another frame. The differential signal between these two frames can be used as the modification signal. Because of the similarity, the frame differential signal usually occupies a very narrow total bandwidth. By using the optimal sampling algorithm developed earlier, a very low sampling rate can be achieved. In addition to the sampling reduction, a new speech scrambling technique based on the multi-band-pass signal sampling algorithm is also proposed. The speech is spectrally scrambled at the transmitter to produce a defined multiple band pass structure. The reconstruction 'key' comprises a knowledge of the spectral frequency bands of the scrambled elements together with a unique sampling rate which will allow correct recombination of these elements.

## 1.4 Organisation of Thesis

This thesis is comprised of seven chapters. Chapter 2 presents an overview of the previous work in the area of low rate transmission of speech. Both speech bandwidth reduction and low bit rate speech coding techniques are investigated, and comparisons are made among the existing techniques.

Chapter 3 describes the new sampling rate reduction algorithms, called

Time Domain Compression technique and Frequency Companding technique, for multi-band-pass signals. These methods are then implemented in speech signals. When a voice signal has a spectrum with suitable gaps, sampling rate reduction algorithms can be employed to compute a minimum sampling rate. When the condition of 'suitable gaps' is not satisfied, the algorithms are modified to incorporate a variable decision threshold. It is shown that, by careful choice of threshold, the sampled signal can be reconstructed without significant loss of quality.

Chapter 4 is a further development of Chapter 3. The optimal sampling rate algorithms developed in Chapter 3 are suitable for multi-band-pass signals. The gaps in the signals spectrum are the key factors which affect the ratio of the sampling rate reduction. This chapter analysis the pseudo-periodic property of voiced speech. A comparison is then made between two consecutive frames and the differences are extracted. Because of the similarity, the differences will occupy a small range of frequency in total. In other words, there are wide gaps among the significant differential components. To sample the differential signal, a very low sampling rate is required. This method, associated with the algorithms developed in Chapter 3 has the capability of Voiced/Unvoiced/Silent classification, pitch detection, differences extraction and speech reconstruction.

The Time Domain Compression algorithm can also be used in speech scrambling. In Chapter 5, a new speech scrambling technique is described. The speech is spectrally scrambled at the transmitter to produce a defined

multiple bandpass structure. The reconstruction 'key' comprises a knowledge of the spectral frequency bands of the scrambled elements together with a unique sampling rate which will allow correct recombination of these elements. To prevent unauthorized reconstruction, the bandpass structure and sampling rate can be changed irregularly according to an appropriate key sequence. Both single user and multi-user systems are developed. Also, the sampling rate tolerance is calculated.

Chapter 6 presents the practical applications and the evaluation of the speech compression techniques. The combination of low sampling rate techniques with low bit rate coding techniques is investigated. Reliable speech transmission using combined adaptive-rate sampling and channel coding techniques is studied. A real-time speech compression, Frequency Companding, using digital signal processor is also presented in this chapter.

Chapter 7 presents a summary of the methods developed and the author's conclusions with respect to the compression applicability. In addition, some areas are suggested for further study using the newly developed compression techniques.

## **Chapter 2**

# **Review of Speech Compression Techniques**



## 2.1 Introduction

Modern speech communications systems are increasingly using digital transmission due to the advantages of digital signal process. Digital communication may be retransmitted many times without significant loss of quality. Some digital methods are extremely tolerant to noise and interference, which means that digital signals can use less power than analog signals for the same transmission range. Security and privacy are easy to maintain. Digital speech can be handled over existing data lines, just like any other digital information.

The fundamental concern for digital speech process is that of representation of speech signals in digital form [Rabiner,*et al.*(1978)]. The analog speech is represented by the samples taken periodically in time; provided that the samples are taken at a high enough rate [Shannon,(1949)].

However, digital speech processing started with a big disadvantage with regard to the bandwidth. For a typical speech bandwidth of 4000 Hz, at sampling rate of 8000 samples per second, and 8 bits to characterize the amplitude of each sample, a total of 64000 bits per second is needed to represent the speech. To reduce the transmission rate, one can either compress the speech bandwidth so that a lower sampling frequency can be used or use low bit rate coding techniques. The former is called analog bandwidth reduction and the latter is called digital bandwidth reduction [Gieseler,*et al.*(1980)]. Over the past decades, the overwhelming majority of attention has been given to dig-

ital rather than analog techniques. Most activity in speech compression is concentrated on digital techniques which provide a bit rate reduction rather than direct audio bandwidth reduction. The primary method used to encode speech signals is Pulse Code Modulation (PCM) [Reeves,(1938)]. PCM is not a bit rate reduction system but a standard bit rate system and the resulting high quality has become the standard of comparison for bit rate reduction systems [Spanias, *et al.*(1992)].

Speech coding techniques can be classified into the following three categories [Holmes,(1988)]:

- waveform coding
- parameter coding
- mid-band coding

Waveform encoding, as its name implies, attempts to copy the actual shape of the waveform produced by the microphone and its analogue circuit. Parameter encoding extracts certain parameters from the original speech and encodes them as digital bits. When the speech is reconstructed, the waveform may not appear to be similar to the original, but will sound similar. The mid-band coding technique is the combination of the waveform coding and various data compression techniques. It normally give better speech reproduction in the 4 - 16 kb/s range than is possible with either of the other two techniques at these digit rates [Holmes,(1988)].

## 2.2 Analog Speech Bandwidth Reduction

### 2.2.1 Frequency Division

Frequency division [Bogner,(1965)] is a technique which compresses the frequencies at the transmitter and expands them at receiver. In this technique, voice bandwidth is divided into three sub-bands, according to the three main vowel formants, which are in the range of 200-1000, 1000-2000 and 2000-3000 Hz respectively. The signal in each channel was then 'frequency divided' by amplitude preserving dividers. This division process is such that the component of largest amplitude (corresponding to the formant) is divided, and other components translated maintaining the original spacing and similarity of amplitudes. Thus each formant could be preserved by using a filter of bandwidth a fraction of the original bandwidth, while some lower amplitude components would be lost. The narrow bands were recombined, transmitted, separated, and then frequency multiplied. Bandwidth would thus be saved by omission of low energy parts of the spectrum. For example, for a speech bandwidth of 4000 Hz, if frequencies divided and multiplied by two, bandwidth could be halved. Thus, the 4000 Hz speech spectrum can be converted to 2000 Hz, transmitted, received and reconverted to 4000 Hz. However, there is a problem in dividing all the frequencies by two, because by doing so, half of the frequency components will be lost. The missing components have to be interpolated in some way at the receiver and it was found that such interpolation can seriously affect the quality of the speech

[Bogner.(1965)] [Malah,*et al.*(1981)].

A variation of frequency compression-expansion technique consists of transforming the real and imaginary components of the FFT data into amplitudes and phases [Wong,*et al.*(1982)] [Patrick,*et al.*(1983)]. One out of two amplitude and phase components is rejected (put to zero) in the case of 2:1 frequency compression. The remaining components are then moved down and the resulting signal is converted back to the time domain by taking IFFT, and transmitted using only half the bandwidth of the original signal. At the receiver, an FFT is taken for the received signal, the amplitude components are moved back to their original position, and the missing components are linearly interpolated. The missing phases are replaced with random values (because of the low correlation of the phase, it can not be interpolated), and the signal is reconverted to the time domain. The reconstructed speech is said to be quite intelligible. However very noticeable distortion and background noise are produced as a result of the incorrect phase associated with the interpolated amplitude components.

### 2.2.2 Time Domain Speech Gapping

This time domain speech gapping [Gieseler,*et al.*(1980)] makes use of the fact that voiced speech is a quasi-periodic signal, a signal which repeats itself in an almost identical period for many times. This is shown in Figure 2.1 for the phrase of 'five members' in the sentence of 'The committee shall only have five members'. The 'i' sound is made up of about 25 nearly identical pitch

periods. The 's' sound is similar to random noise of no particular pattern. If these signals are chopped into properly-sized intervals: half of the intervals discarded and the remaining transmitted, at the receiver the speech can be reconstructed by simply repeating the received intervals to fill the absent gaps. This technique is often explained by numbering the speech signal intervals 1,2,3,4... . If every other interval is chopped out, the transmitted signal is the sequence of 1,3,5... . At the receiver, each received interval is repeated once to make up the deleted interval. The reconstructed signal is, therefore, the sequence of 1,1,3,3,5,5... . Basically this technique creates time gaps - the transmitted signal is absent half of time. These time gaps can be translated into bandwidth saving by stretching out the signal in each of the transmitted intervals so that it covers the gaps created. A saving of one-half in the time domain can be translated into a bandwidth saving of one-half. Alternatively a second channel can be time division multiplexed into the gaps so allowing two speech signals to occupy the bandwidth normally required for one signal.

A variation of the this technique was proposed by Jibbe [Jibbe,(1986)]. In this technique, the input speech is first classified into one of the patterns - Voiced speech, Unvoiced speech and Silence (VUS) - by a VUS detector. Different transmission scheme is used for different pattern. In a voiced segment, the pitch period is detected and the pitch interval repetition is counted. Only one complete pitch period speech is transmitted together with side information which include VUS classification, pitch period and number of pitch

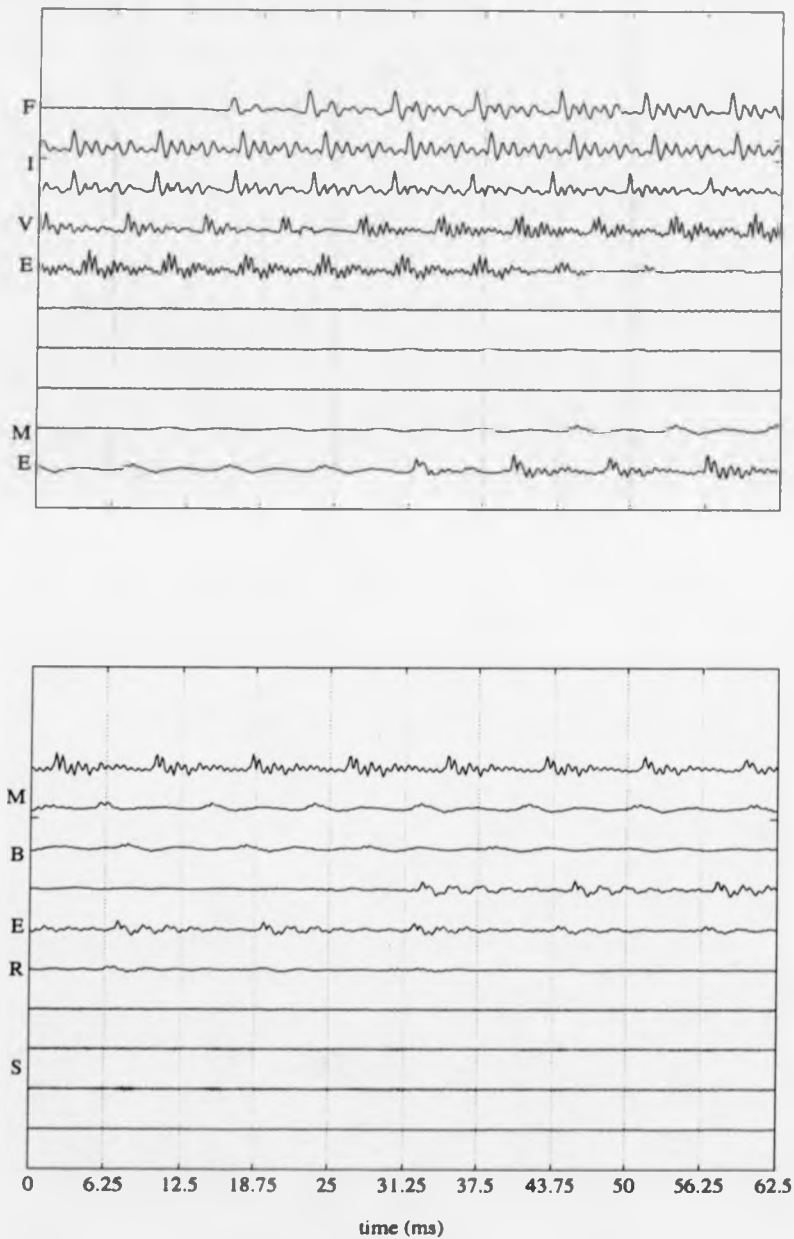


Figure 2.1: Speech waveform for the phrase 'five members'.

repetitions. For an unvoiced frame, the repetition count is always unity indicating no repetition, so in this case the complete frame is transmitted. For a silence frame, only the VUS classification is transmitted, effectively meaning the elimination of a long pause in speech.

The reconstructed speech from the above techniques is intelligible but sounds noticeably artificial resulting from the gapping.

## 2.3 Waveform Encoding Techniques

### 2.3.1 Pulse Code Modulation (PCM)

PCM is the most common method of digital encoding system for speech [Reeves,(1938)]. It is now widely used for feeding analog speech signal into computer or other digital equipment for further process (in this case it is called Analogue-to-Digital (A/D) Conversion).

Figure 2.2 shows a simplified block diagram of a PCM system. In this system, the analogue speech signal is sampled at a frequency of  $\rho_s$  (typically  $\rho_s = 8000$  Hz). The PCM encoder encodes the quantised samples into binary-coded digits which are transmitted over a communication channel. At the receiver, the pulses are detected and re-shaped. These reconstructed pulses are then decoded into the amplitude of the samples which are subsequently processed and filtered to generate an estimate of the original analogue speech waveform. The quantisation noise of simple PCM is determined by the step size associated with a unit increment of the binary

code [Jayant,(1974)]. During low-level speech or silence this noise might be very noticeable, whereas during loud speech it would be masked by the wanted signal. For a given subjective degradation in PCM it is possible to allow the quantisation noise to vary with signal level, so a quality improvement can be obtained [Cutler,(1952)]. This variation can be achieved either by using a non-uniform distribution of quantisation levels or by making the quantisation step size change as the short-term average speech level varies.

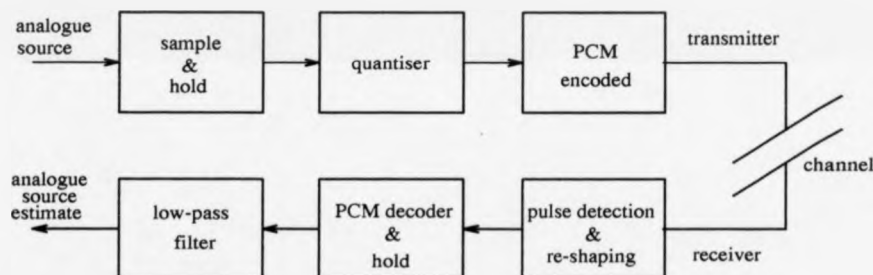


Figure 2.2: Simplified block diagram of a PCM system

The logarithmic quantiser [Smith,(1957)] is the most popular non-uniform quantising level method which allows a wider dynamic range and more accurate waveform representation at lower amplitude levels. Conventionally, the amplitude compression characteristic achieved by a logarithmic type of processing is referred to as 'companding'. For toll quality, a bit rate of 64 kb/s is necessary. For slightly poorer quality, the rate can be reduced to 48 kb/s.



### 2.3.2 Differential Pulse Code Modulation (DPCM)

Various modifications have been made to the basic PCM principle in an attempt to reduce the transmission rate required without reducing the speech quality significantly. The best known of these is differentially encoded PCM (DPCM) [Cutler,(1952)]. DPCM takes advantage of the fact that speech sampled at the Nyquist rate exhibits a very significant correlation between successive samples. This high correlation relation makes it possible to accurately predict a sample by the previous sample value. In simple DPCM only the difference between adjacent samples is encoded and transmitted [Max,(1960)] [Paez *et al.*(1972)]. In more sophisticated DPCM systems, a sequence of past sample values is used to form the estimate of the current sample value and the difference between this estimate and the actual sample value is encoded and transmitted [Jayant,(1974)]. This processing technique results in a saving of about one bit per sample, giving an overall data rate of about 56 kb/s.

Practical DPCM systems are Adaptive DPCM (ADPCM) [Cummsiskey,*et al.*(1973)] in which the quantisation levels or step sizes adjust or adapt to the input signal level. ADPCM systems are always better than PCM for speech signal [Rabiner,*et al.*(1978)]. ADPCM systems generally operating at bit rates from 32 to 40 kb/s give speech quality which most observers find identical to 64 kb/s PCM.

### 2.3.3 Delta Modulation (DM)

Delta modulation [DeJager,(1952)] is actually an ADPCM system in which the difference between adjacent samples values is encoded into a one-bit word. The bit rate is therefore equal to the sampling rate. To compensate for this reduction, the sampling rate must be many times higher than the Nyquist rate so that the adjacent samples become highly correlated, ensuring that the system output will not be dominated by quantisation noise [Jayant,(1974)].

A block diagram of a basic DM system is shown in Figure 2.3.

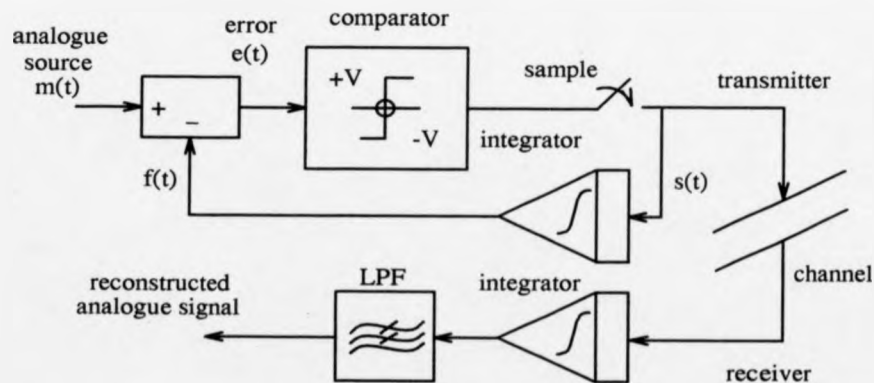


Figure 2.3: Simplified block diagram of a DM system

The transmitter comprises a comparator whose input is the difference between the analogue signal to be transmitted ( $m(t)$ ) and a sampled and integrated version of its output signal ( $f(t)$ ). Thus, when

$$m(t) - f(t) > 0 \quad (2.1)$$

a positive pulse of amplitude  $+V$  and width  $\Delta t$  is generated. Similarly, when

$$m(t) - f(t) < 0 \quad (2.2)$$

a negative pulse of amplitude  $-V$  and width  $\Delta t$  is generated. The integrated version of the sampled comparator output,  $f(t)$ , effectively tracks the input waveform. At the receiver,  $f(t)$  can be recovered by simply integrating the received signal; low-pass filtering, and then removes the higher frequency components associated with the quantised signal  $f(t)$  to yield an estimate of the speech signal  $m(t)$ .

Delta modulation techniques can only operate over a limited input signal dynamic range before the signal-to-quantisation noise ratio at the output become excessive. Practical systems have an adaptive step size similar to ADPCM and are designated as Adaptive Delta Modulation (ADM) [Abate,(1967)]. ADM systems generally operate at bit rates from 16 to 32 kb/s. At 32 kb/s the quantisation noise is virtually inaudible. At 16 kb/s the noise is noticeable but does not impair intelligibility [Gieseler,*et al.*(1980)].

#### 2.3.4 Sub-Band Encoding (SBC)

Sub-band coding of speech [Crochiere,(1977)] is a waveform coding technique which makes an optimum bits allocation to different part of speech signal. The speech frequency is first divided into many sub-bands (typically between

4 and 8) by a bank of bandpass filters. Each sub-band is then sampled at its Nyquist rate (twice the width of the band) and digitally encoded (Figure 2.4). The principle for the coder is that the bit allocation can be weighted so that those sub-bands with the most important information get most bits. The initial sub-band encoding technique used fix bit allocations based on the average spectrum of speech [Crochiere.*et al.*(1982)]. A variant version introduced an idea of dynamically changing the bit allocation [Ramstad.(1982)]. This idea is to quantise and transmit the root mean square (rms) value of each of the bands for a frame of speech. Based on the quantised value, the remaining bits could be allocated among the sub-bands in an optimal manner. This modified version of SBC has been developed into both time domain and frequency domain [Honda.*et al.*(1985)] [Soong.*et al.*(1986)]. The sub-band coding technique provides toll quality speech in the region from 16 to 24 kb/s [Holmes.(1988)].

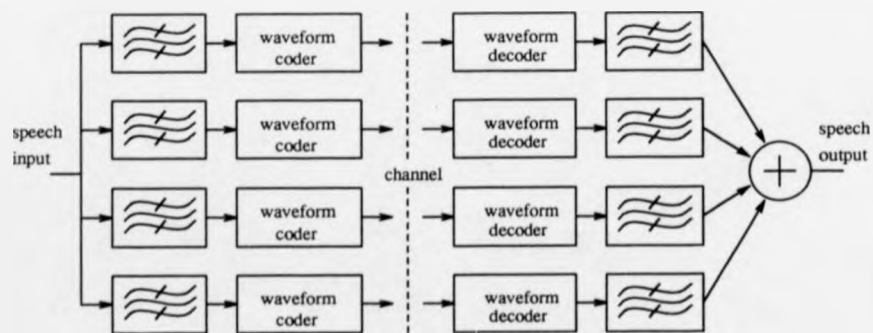


Figure 2.4: Block diagram of sub-band encoding system

### 2.3.5 Transform Coding (TC)

Orthogonal transformations of speech signals provide a possibility of bit rate reduction for transmission of signals. A mathematical transformation, such as Fourier [Robinson,*et al.*(1970)] [Campanella,*et al.*(1971.a)] or Walsh-Hadamard [Campanella,*et al.*(1971.a)] [Robinson,(1971)] is first applied to segments of a speech signal. Then the transform coefficients are quantised and transmitted. The primary advantage of this technique is that the system can be adaptive so that transform coefficients which are unimportant can be omitted and not transmitted at all.

One of the earliest applications of orthogonal transformations to the processing of speech signals was carried out by Boesswetter [Boesswetter,(1970)]. In his work, a 1.4 second of a speech signal was examined. The speech was sampled at 8000 Hz and then divided into short time frames. Each frame contained 16 samples, equivalent to 2 ms. The results showed that if the 16 samples were transformed using a 16-point Hadamard Transformation, reasonably intelligible speech could be reconstructed by using only the two largest coefficients from the set of 16 Hadamard coefficients. Campanella and Robinson [Campanella,*et al.*(1971.b)] have shown mathematically that, by using a 16-point Fourier and Hadamard transformations, it should be possible, for a given signal-to-quantisation noise ratio, to transmit speech using 46 kb/s for Fourier transformation and 48.5 kb/s for Hadamard transformation, rather than 56 kb/s as in conventional PCM.

## 2.4 Parameter Encoding Technique

In waveform encoding techniques, the coding systems try to faithfully reproduce the input waveform. However parameter coding systems make use one of the basic properties of speech production that gives a great saving in digit rate. As it was mentioned earlier, human speech can be classified into three categories, voiced, unvoiced and silence. Analysed over a short time frame, a few tens of milliseconds, the glottal excitation for voiced speech have two distinct properties: (i) it is 'periodic' and (ii) its variation within a period is smooth. The excitation for unvoiced speech on the other hand is considered to be random white noise. The excitation mode of speech production is shown in Figure 2.5. In the frequency domain, the fine structure of the spectrum consists either of lines, resulting from harmonic of the fundamental frequency, or is continuous, during random noise excitation. The envelope of the speech spectrum is the result of combining the spectral trend of the corresponding sound source with the response of the vocal tract. By separating the fine structure specification of the sound sources from the overall spectral envelope description, and specifying both in terms of a fairly small number of slowly varying parameters, it is possible to transmit a reasonably adequate description of the speech at data rates of 1000 - 3000 b/s. At the receiver, the speech is resynthesised by using either a periodic or random noise source feeding a dynamically controlled spectral shaping filter. The basic parameter encoding system is shown in Figure 2.6.

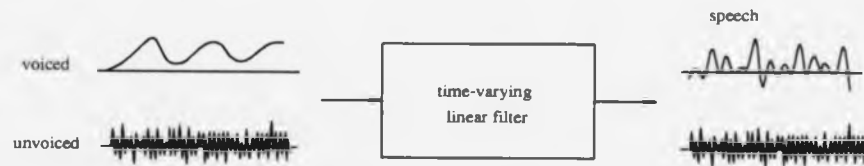


Figure 2.5: Excitation model of speech generation

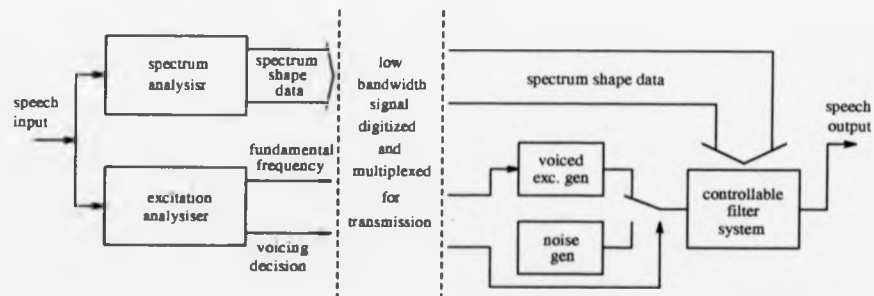


Figure 2.6: Block diagram of the basic elements of a parameter coding system

The three most important types of parameter coding techniques are known as linear predictive coding [Makhoul,(1975)], channel vocoders [Kelly,(1970)] and formant vocoder [Rosenberg,*et al.*(1971)].

### 2.4.1 Voiced/Unvoiced/Silence and Pitch Detection

A pitch detector is an essential component in a variety of speech processing systems. Besides providing valuable insights into the nature of the excitation source for speech production, the pitch contour of an utterance is useful for speech recognition [Atal, *et al.*(1976)], for speech instruction to the hearing impaired [Levitt,(1973)], and is required in almost all speech analysis synthesis (vocoder) systems [Flanagan,(1972)]. Because of the importance of pitch detection, a wide variety of algorithms for pitch detection have been proposed. The following are some of the most important algorithms for pitch detection.

#### (i) Modified Autocorrelation Method Using Clipping

The modified autocorrelation pitch detector [Dubnowsk,*et al.*(1976)] is based on the centre-clipping method. This method requires calculating the autocorrelation function of a speech signal filtered to 900 Hz. The autocorrelation function is clipped at a certain level (computed in the algorithm), and is searched for its normalised maximum value. If the normalised maximum value exceeds 0.3, the section is classified as voiced and the location of the maximum is the pitch period. Otherwise, the section is classified as unvoiced.



In addition to the voiced/unvoiced classification based on the autocorrelation function, a preliminary test is carried out on each frame of speech to determine if the peak signal amplitude within the section is sufficiently large to warrant the pitch computation. If the peak signal level within the frame is below a given threshold the frame is classified as unvoiced (silence) and no pitch computations are made.

#### (ii) Cepstrum Method

A spectrum of the logarithm of a frequency spectrum is defined as cepstrum [Bogert.(1963)]. The cepstrum pitch detector starts as follows [Schafe,*et al.*(1970)]: each frame of 512 samples is weighted by a 512-point Hamming window, and then the cepstrum of the frame is computed. The peak cepstral value and its location is determined and if the value exceeds a fixed threshold, the frame is classified as voiced speech and the pitch period is the location of the peak. If the peak does not exceed the threshold, a zero-crossing count is made on the block. If the zero-crossing count exceeds a given threshold, the frame is called unvoiced. Otherwise, it is called voiced speech and the location of the maximum value of the cepstrum is the pitch period. A preliminary silence detector is used to classify all low-level block as silence prior to the cepstrum computation.

**(iii) Simplified Inverse Filtering Technique (SIFT)**

The SIFT method of pitch detection starts as follows [Markel,(1972)]: a block of 400 samples taken at a rate of 10000 Hz is low-pass filtered to a bandwidth of 900 Hz, and then decimated (down sampled) by a 5 to 1 ratio. The coefficients of a 4th-order inverse filter are obtained by using the autocorrelation method. The 2000 Hz speech signal is then inverse filtered to give a spectrally flattened signal which is then autocorrelated. The pitch period is obtained by interpolating the autocorrelation function. A voiced/unvoiced decision is made on the basis of the peak of the autocorrelation function.

The threshold used for this test is normalized value of 0.4 for the autocorrelation peak. As with the previous two pitch detectors, a preliminary silence detector is used to classify low-level sections as silence and eliminate them from further consideration.

**(iv) Data Reduction Method**

This pitch detector [Miller,(1975)] places pitch markers directly on a low-pass filtered (0-900 Hz) speech signal and thus is a pitch synchronous pitch detector. To obtain the appropriate pitch markers, the data reduction method first detects excursion cycles in the waveform based on intervals between major zero crossings.

The remainder of the algorithm tries to isolate and identify the principal excursion cycles, i.e., those which correspond to true pitch periods. This is accomplished through a series of steps using energy measurements and logic

based on permissible pitch periods and anticipated syllabic rate changes of pitch. An error correction is used to provide a reasonable measure of continuity in the pitch markers. Since there is no inherent voiced/unvoiced calculation within this pitch detector, regions of unvoiced speech are identified by the lack of pitch markers.

#### **(v) Parallel Processing Method**

The parallel processing pitch detector [Gold, *et al.* (1969)] receives the low-pass filtered (900 Hz) speech signal. Then a series of measurements are made on the peaks and valleys of the low-pass filtered signal to give six separate functions. Each of these six functions is processed by an elementary pitch period estimator, giving six separate estimates of the pitch period. The six pitch estimates are then combined by a sophisticated decision algorithm which determines the pitch period. A voiced/unvoiced decision is obtained based on the degree of agreement among the six pitch detectors. Additionally, the preliminary silent detector is used to classify low-pass segment as silence.

#### **(vi) Average Magnitude Difference Function (AMDF)**

The AMDF pitch detector [Ross, *et al.* (1974)] starts by initially sampling the speech signal at 10000 Hz. A zero-crossing measurement is made on a the full-band speech file, and an energy measurement is made on a low-pass filtered version (0-900 Hz) of the signal. The average magnitude difference function is computed on the low-pass filtered speech signal at 48 lags run-

ning from 16 to 126 samples. The pitch period is identified as the value of the lag at which the minimum AMDF occurs. Thus a fairly coarse quantization is obtained for the pitch period. Logic is used to check for pitch period doubling, etc., and to check on continuity of pitch period with previous pitch estimates. In addition to the pitch estimates, the ratio between the maximum and minimum values of AMDF is obtained. This measurement, along with zero-crossing measurement and energy measurement is used to make a voiced/unvoiced decision using logical operation.

#### 2.4.2 Linear Predictive Coding (LPC)

In the LPC technique, the vocal tract model is based on the principle that speech can be reasonably predicted by weighting the sum of previous speech samples [Rabiner, *et al.* (1978)]. Models of speech production process usually treat the vocal tract and the air entering the vocal tract (the excitation) separately. In LPC analysis [Makhoul, (1975)] the vocal tract section of the speech production model, Figure 2.7, is represented by a time-varying linear digital filter. This filter must represent the effects of lip radiation, glottal pulse shape and nasal cavity. The aim of LPC analysis is to extract the set of parameters from the speech signal which specifies the filter transfer function which gives the best match to the speech to be coded. An all-pole filter of order  $p$  (usually in the range 10 to 20) is used to model the vocal tract.

Consider the LPC all-pole model of the vocal tract, the transfer function

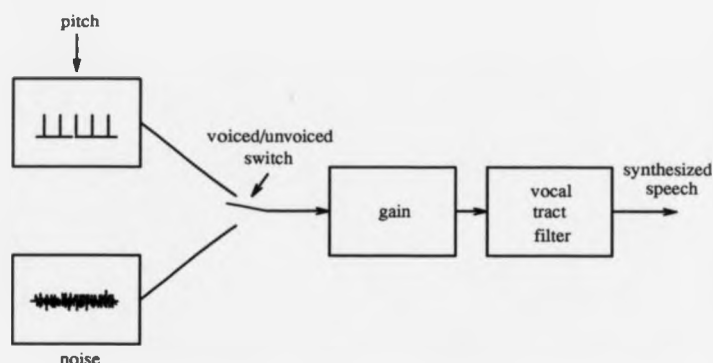


Figure 2.7: Speech production model.

of the filter has the form of

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (2.3)$$

and the parameters of the LPC synthesis model are (i) the voiced/unvoiced indication and the pitch period, (ii) the gain parameter  $G$  and (iii) the filter coefficients ( $a_k$ ). This simplified all-pole model is a natural representation of non-nasal sounds, but for nasals and fricative sounds both poles and zeros are required in the vocal tract transfer function. If however, the order of the filter  $p$  is large enough, the all-pole model is a good approximation for almost all the speech sounds.

Having made the assumption that an all-pole filter of order  $p$  will model the vocal tract, the LPC analysis must extract sets of coefficients ( $a_k$ ) periodically from the speech signal. The vocal tract shape generally changes relatively slowly and it is thus sufficient to update the parameters once every

10 to 20 ms.

The output samples  $s(n)$  are related to the excitation  $u(n)$  by the simple difference equation

$$s(n) = \sum_{k=1}^p a_k s(n-k) + Gu(n) \quad (2.4)$$

The signal  $s(n)$  can be predicted approximately from a linearly weighted summation of past samples. Let this approximation of  $s(n)$  be  $\bar{s}(n)$ , where

$$\bar{s}(n) = \sum_{k=1}^p a_k s(n-k) \quad (2.5)$$

The error between the actual value  $s(n)$  and the predicted value  $\bar{s}(n)$  is given by

$$e(n) = s(n) - \bar{s}(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (2.6)$$

$e(n)$  is also known as the residual. The prediction coefficients ( $a_k$ ) are calculated to minimize the prediction error. So the predicted characteristics obtained can be used to resynthesise the corresponding frame at the receiver. The two most important methods to be used for optimisation are known as autocorrelation [Markel, *et al.*(1973)] and covariance methods [Atal, *et al.*(1971)].

### 2.4.3 Channel Vocoders

The basic idea of the channel vocoder is to attempt direct representation of the combined vocal tract and excitation power spectrum by summing together suitably weighted fixed frequency responses [Kelly,(1970)]. The block diagram of the channel vocoders system is shown in Figure 2.8.

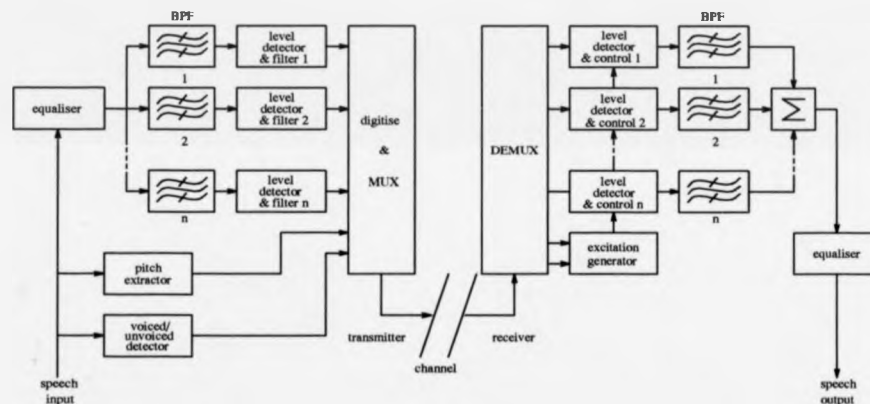


Figure 2.8: Simplified block diagram of a channel vocoder.

A channel vocoder analyses the baseband speech signal by applying it to a bank of  $n$  contiguous band-pass filters and then detecting and encoding the output levels from these filters. Unless a very large number of channels can be used (with consequent high digit rate) it is difficult to achieve a good approximation to the spectrum shapes around the formant peaks with a channel vocoder. However, the quality achievable with around 15 - 20 channels is reasonably acceptable for communication purposes, and does not require too high a data rate for transmission [Holmes,(1988)]. The analogue speech in-

put is first equalised to ensure that the input to all the analysis BPF's have similar average levels over the complete baseband frequency range. Logarithmic processing and detection are applied to each of the filter outputs to give a set of relatively slowly varying signals representing the band-pass spectral envelopes; these signals are then sampled and digitised. Similar to that in LPC, the input speech is analysed to determine whether the sound is voiced or unvoiced. If it is voiced sound, the pitch is detected. A complete data frame comprising regularly sampled values of filter output levels, voiced/unvoiced indication, pitch frequency and overall amplitude information is then transmitted in multiplex form at the vocoder frame rate.

At the receiver, the data frames are demultiplexed and decoded. After antilog processing, the decoded filter outputs are used to control the gain of the synthesis filters. Again, in a manner similar to that described previously for LPC, the synthesis filters are excited by a noise-like signal for unvoiced speech and by a pitch frequency generator for voiced speech.

Channel vocoding is a robust speech digitisation technique, of which there are many practical variants. Channel vocoders typically operate in the range 1200 b/s to 9600 b/s with roughly 600 b/s are used to the pitch and voicing information and the remaining information devoted to the channel signals [Rabiner, *et al.* (1978)].



#### 2.4.4 Formant Vocoders

In general, for a normal speaker, speech energy will be concentrated in a few well defined regions of the baseband spectrum; these regions are centred around the so-called formant frequencies which will clearly vary from speaker to speaker. Formant vocoders [Rosenberg, *et al.*(1971)] use a synthesiser that is much more closely related to human speech production, because not only they use the choice of periodic or noise sources as in LPC and channel vocoders, but also the spectral filter system has resonators that are explicitly related to the principal formants of the input speech. Therefore, the formant vocoders concentrate on modeling speech specifically in the vicinity of the formants by describing accurately the energy distribution in these regions as a function of time. A simplified block diagram of formant vocoders is shown in Figure 2.9.

Here, the speech input is analysed in terms of sampled values of five formant frequencies together with the energy levels associated with each formant region. Pitch analysis and voiced/unvoiced detection are carried out as for the LPC and channel vocoders. This array of data is sampled, digitised and multiplexed prior to transmission. At the receiver, the decoded formant frequencies and amplitudes are used to set the parameters of adjustable resonant circuits to appropriate values; these circuits are then excited with a pitch frequency or noise-like signal according to whether the sound is voiced or unvoiced in a particular frame interval.

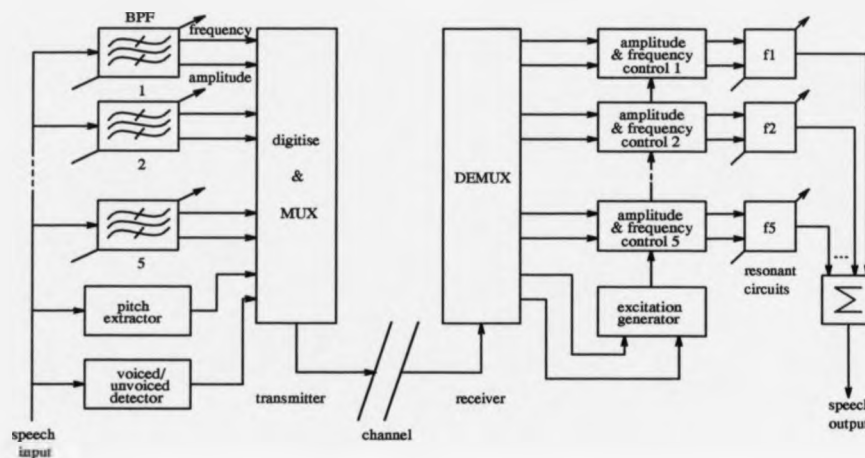


Figure 2.9: Simplified block diagram of a formant vocoder

Formant vocoders have been demonstrated to work effectively down to rates of 1.2 kb/s or less; they can, however, be sensitive to utterances which do not have a well-defined formant structure. In principle, the formant vocoder represents a more sophisticated model of the speech process than either LPC or channel vocoding and therefore would appear to have the potential for lower transmission rates than either of these.

## 2.5 Mid-Band Encoding Techniques

In waveform encoding systems, the typical transmission rate is in the range of 16 - 64 kb/s. Since the conventional analogue speech communication bandwidth is about 3000 to 4000 Hz, waveform encoding represents a bandwidth expansion process since a bandwidth considerably in excess of the analogue

speech bandwidth will normally be required for the transmission of the corresponding digitised data. With many practical communications media, efficient bandwidth utilisation is an important consideration and thus any significant reduction in speech digitisation rate, and hence occupied bandwidth, is beneficial.

Several techniques which attempt to digitise speech in the so-called 'mid-band' between about 4.8 and 16 kb/s were developed. Mid-band systems can be considered as the bridge between waveform and parameter encoding techniques, they fall into following three categories [Holmes,(1988)]:

### **2.5.1 Combine Waveform Encoding with Data Compression**

As examples of this category, References [Turner,(1976)] [Frangoulis,(1977)] describe two methods of data compression for speech signals. In [Turner,(1976)], an adaptive 9.6 kb/s, 2-bit PCM system based upon that proposed by Wilkinson [Wilkinson,(1973)] is described. The principle of this system is shown in Figure 2.10. At each sampling instant, the 2-bit samples are made up of one bit to indicate the polarity of the sample and one bit to indicate whether the sample value was inside or outside the boundary set by the threshold levels. The threshold levels were varied adaptively in accordance with the short-term average level of the speech signal. However, it was found that the basic 9.6 kb/s data contained so little redundancy that significant data

compression factors were not achievable.

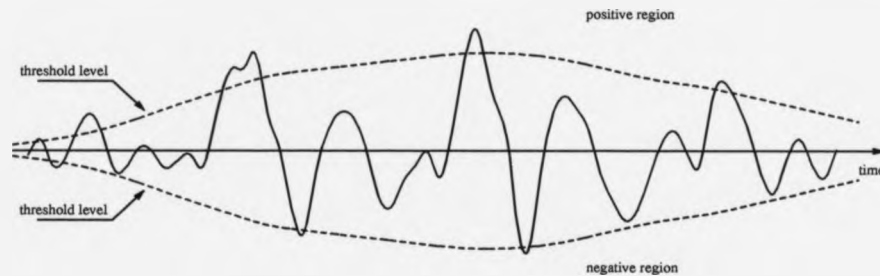


Figure 2.10: Waveform for adaptive 2-bit PCM system

An application of the Hadamard transform to the compression of speech bit rates was described in [Frangoulis,(1977)]. Here, the digitised samples of speech are analysed in terms of their Hadamard transform coefficients. Only the dominant coefficients are then transmitted over the communications channel in quantised form; at the receiver, the inverse Hadamard transformation is performed in order to reconstruct an estimate of the original speech signal. This technique promises to provide a reasonable quality speech at a bit rate about 8 kb/s [Frangoulis,(1977)].

### 2.5.2 Advanced Waveform Encoding

An example of this category of mid-band encoding techniques is described in [Alexander,(1979)]. The technique, known as time encoding of digital speech, encodes the input waveform in terms of parameters such as (a) lengths of time between zero crossings of the speech waveform, (b) the number of local

maxima and minima in speech segments between zero crossings and (c) amplitude information. These parameters enable each segment of the waveform between successive zero crossings to be characterised by one template. At the receiver, using an identical library of templates, the encoded parameter information is decoded in order to estimate the profile of the input speech signal; clearly, in the absence of noise, this sequence will be identical with that computed at the transmitter.

### 2.5.3 Simplified Parameter Encoding Schemes

The most important technique in this category is adaptive predictive coding (APC) [Sambur,(1982)]. APC is a variant of LPC and can also be seen as a form of differential encoding. In LPC, the vocal tract is represented in the form of a mathematical model. By employing a linear combination of past values of the speech samples and the LPC filter coefficients for the current frame, it is possible to predict the proceeding speech samples. The data which are transmitted in the LPC system are (a) LPC coefficients  $a_i$ , ( $i = 1, \dots, p$ , where  $p$  is the number of poles), (b) pitch period  $T$ , (c) the gain  $G$  and (d) the voiced/unvoiced indication. In general, the APC transmitter is configured as is shown in Figure 2.11.

The coefficients  $a_i$  are calculated to minimise the error between a quantised version of the speech input and the speech signal produced by LPC analysis/synthesis. In LPC systems, the error signal is not transmitted whereas in APC it is. However, if the LPC algorithm is efficient, the amount of

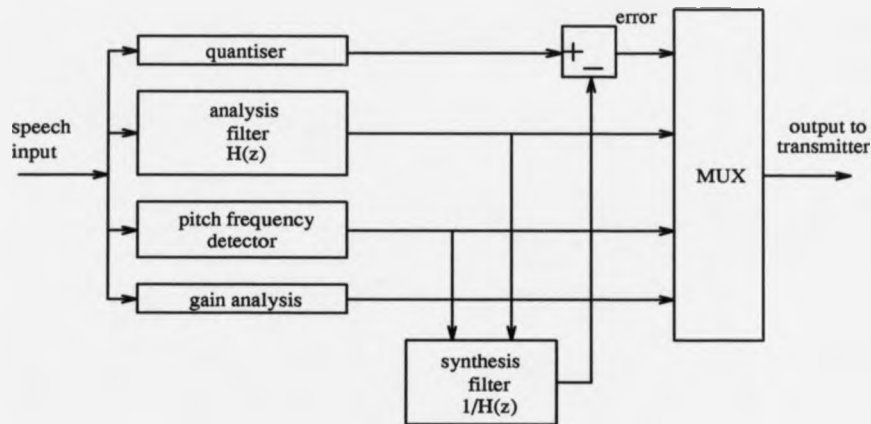


Figure 2.11: Simplified block diagram of APC transmitter

information in the error signal will be relatively small.

In practical APC systems, the number of poles,  $p$ , is typically 4 and the overall transmission rate approximately 9.6 kb/s.

## 2.6 Comparison of Speech Coding Techniques

Generally speaking, waveform coding requires a high bit rate ( $> 16$  kb/s), but can give toll quality reconstructed speech; vocoding requires a very low bit rate ( $< 4.8$  kb/s) but gives only synthetic quality speech and mid-band coding requires a low bit rate (4.8 – 16 kb/s) and gives speech quality that is not quite as good as that of waveform coders. Comparisons of the bit rates of various coding techniques are shown in Figure 2.12. At the high quality end of the scale PCM is used as a standard for low bit rate speech coding

techniques to be compared with. PCM does not exploit any of the special properties of speech production or auditory perception. It encodes all other signals equally well. The other coders listed in the figure are designed for speech signals and do not operate well with other types of signals.

The ADPCM operate well in the region above 32 kb/s and, at about 40 kb/s provide toll quality [Jayant,(1974)]. ADM has found use in multi-channel rural carrier systems in telephony and in certain satellite communication systems. The SBC and ATC operate best in the region from about 12 kb/s to 24 kb/s. At 12 kb/s they tend to be noisy and somewhat synthetic. About 24 kb/s their quality is quite good but simpler coders such as ADM and ADPCM can provide roughly the same quality. Below 12 kb/s, only synthetic quality speech can be obtained by using parameter coding.

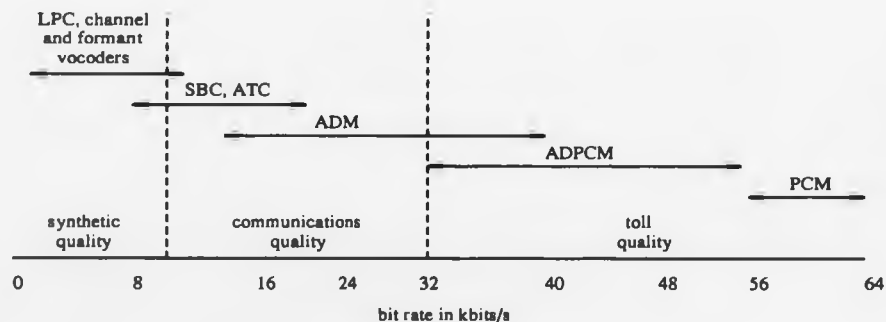


Figure 2.12: Bit rates of speech encoders

All the results presented in Figure 2.12 are for the speech coding techniques operated on the speech samples taken at the Nyquist rate. The transmission rate can be reduced further if lower sampling rate can be applied.

## Chapter 3

# Adaptive-Rate Sampling for Speech Compression



### 3.1 Introduction

Conventional speech compression via speech coding attempts to make more efficient use of communication channels by reducing the inherent redundancy in speech sources to give a lower transmission rate. The speech coding techniques either try to find an optimum code allocation to different parts of signals [Turner,(1976)] [Ramstad,(1982)] [Honda,*et al.*(1985)] [Soong,*et al.*(1986)] or extract and encode the speech parameters [Kelly,(1970)] [Makhoul,(1975)] to achieve a bit rate reduction. However, less attention has been given to the more direct approach of determining the sampling rate. Clearly, the more efficient the sampling processing, the greater is the potential efficiency of the overall digital transmission process. Efficient sampling is also advantageous when it is necessary to store sampled analogue data in a long-term storage medium, e.g. magnetic tape, or where it is required to transmit the digital samples over a communication channel having a restricted information transmission capacity.

In general, the conventional approach to determine the sampling rate for a given analogue signal is based on the Nyquist criterion which will be briefly reviewed in next section. This relates the sampling rate to the maximum frequency, or to the difference of the maximum and minimum frequencies of the analogue signal to be sampled. In speech communication, the voice bandwidth is typically defined to be in the range between 300 and 3400 Hz. In order to obtain a good quality reconstructed speech, the sampling rate

must be at least  $2 \times 3400 = 6800/s$ , or say  $8000/s$  to allow for practical constraints. However, it has been shown that improved sampling rates are possible for multiple band-pass signals where the bands are reasonably distinct [Dodson, *et al.*(1985)] [Dodson, *et al.*(1988)].

This chapter first reviews the conventional sampling theorem, the Shannon sampling theorem. Then we will consider the procedures for spectral manipulation, applied to signals having spectra comprising multiple ( $\geq 2$ ) separated band-pass elements. In essence, the algorithm provides an analytical and practical tool to enable spectral manipulation to be carried out in a systematic manner. This allows non-uniform, adaptive sampling rates to be applied to a multiple band-pass analogue signal if the band structure is first characterised.

By analysis of the form of a spectrum comprising of multiple band-pass elements, the procedure can be used to derive the optimum (i.e. the minimum) sampling rate that is necessary to characterise the spectrum completely, and hence allows its complete recovery from the samples taken at that calculated rate.

These procedures can be applied to the digitisation of signals of various types in order to minimise the number of samples per unit time which must be transmitted over a communication channel, or stored in a storage medium, to enable the data to be completely recovered after transmission or storage. These procedures can therefore be viewed as a form of data minimisation or data compression with respect to the classical Nyquist criterion.

### 3.2 Shannon Sampling Theorem

The simplest case is that of a time function  $f(t)$  whose spectrum  $F(\omega)$  is limited to  $-W \leq \omega \leq W$  (Figure 3.1). If the samples are taken at the regular intervals spaced  $\tau$  apart, the sampled signal denoted by  $\tilde{f}(t)$  is given by

$$\tilde{f}(t) = f(t) \sum_n \delta(t - n\tau) = \sum_n f(n\tau) \delta(t - n\tau) \quad (3.1)$$

The transform of  $\sum_n \delta(t - n\tau)$  is  $\sum_n \delta(\omega - n/\tau)/\tau$  [Linden,(1959)].

The equivalent of time domain multiplication of waveforms is the convolution of corresponding spectra and the equation (3.1) leads to

$$\tilde{F}(\omega) = F(\omega) * \sum_n \frac{1}{\tau} \delta(\omega - \frac{n}{\tau}) = \sum_n \frac{1}{\tau} F(\omega - \frac{n}{\tau}) \quad (3.2)$$

Apart from the weighting factor  $1/\tau$ ,  $\tilde{F}(\omega)$  is seen to consist of replicas of  $F(\omega)$  centered at  $\delta(\omega - n/\tau)$  as shown in Figure 3.1

Figure 3.1(b) shows that if the sampling period is too large, in other words, if the sampling rate is too low, the shifted versions of  $F(\omega)$  overlap, and the original signal cannot be reconstructed from the samples without loss of information. If the sampling rate is  $1/\tau = 2W$ , Figure 3.1(c), or higher, Figure 3.1(d), the overlap will not happen. The original spectrum may be recovered by multiplying  $F(\omega)$  by the spectral window function  $S(\omega)$  shown in Figure 3.1(e). The equivalent operation in the time domain is the

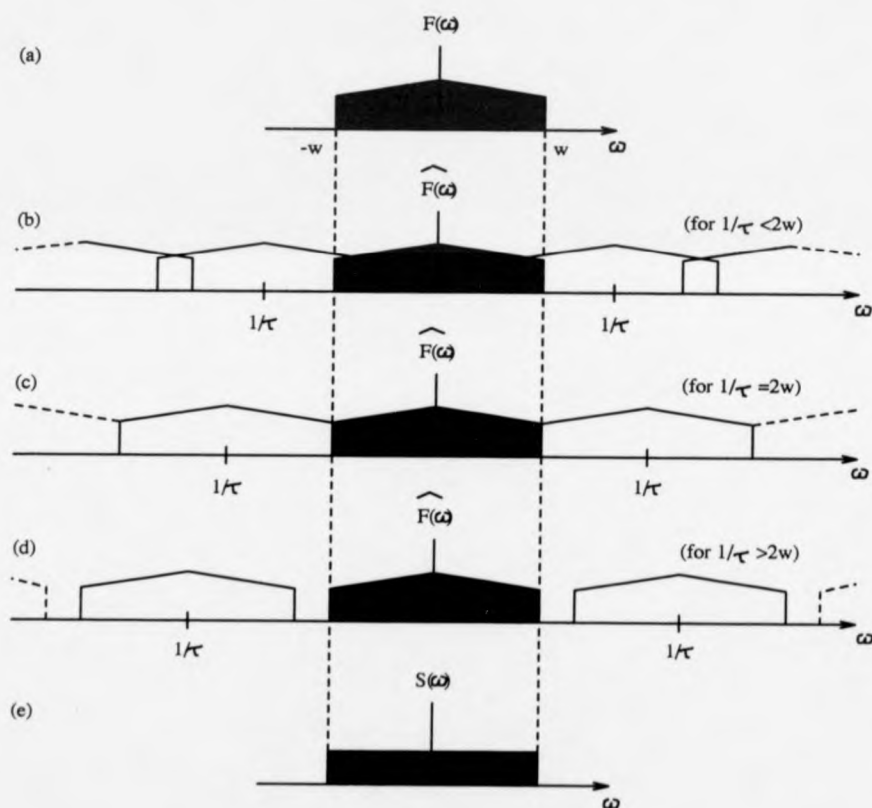


Figure 3.1: Spectrum of sampled low pass signal

convolution of  $f(t)$  by the inverse Fourier transform  $s(t)$  of  $S(\omega)$ , i.e.

$$f(t) = s(t) * \sum_n f(n\tau)\delta(t - n\tau) = \sum_n f(n\tau)s(t - n\tau) \quad (3.3)$$

Substituting  $\tau = 1/2W$  and the functional form of  $s(t)$

$$f(t) = \sum_n f\left(\frac{n}{2W}\right) \frac{\sin 2\pi W(t - n/2W)}{2\pi W(t - n/2W)} \quad (3.4)$$

The original Shannon statement [Shannon,(1949)] of the sampling theorem is as follows

*If a function  $f(t)$  contains no frequencies higher than  $W$  cps it is completely determined by giving its ordinates at a series of points spaced  $(1/2W)s$  apart.*

It is useful to think of the sampling frequency or rate (or the reciprocal of the sampling interval) as a fundamental length in the frequency domain (or reciprocal space); the spectrum and its translates by the sampling rate repeat periodically without overlapping along the frequency axis with a period equal to the sampling rate.

A similar but slightly more complicated construction lies at the heart of the determination of the sampling rate for a bandpass signal, with frequencies confined to an interval  $(\omega_L, \omega_H)$ . In this case it can be verified that translates of the spectrum by integer multiples of  $2\omega_H / [\omega_H / (\omega_H - \omega_L)]$  ( $\geq 2(\omega_H - \omega_L)$ ) also do not overlap. ( $[A]$  is the integer part of the real number  $A$ ). Thus

$2\omega_H / \lfloor \omega_H / (\omega_H - \omega_L) \rfloor$  is a fundamental length for the high bandpass signal (Figure 3.2). Note that a translate of the reflected or reversed band at  $(-\omega_H, -\omega_L)$  also appears in the fundamental interval  $(0, 2\omega_H / \lfloor \omega_H / (\omega_H - \omega_L) \rfloor)$ . When the signal has a high frequency carrier wave modulated by audio or low frequencies, the modulating frequencies can be obtained by sampling at a rate  $2\omega_H / \lfloor \omega_H / (\omega_H - \omega_L) \rfloor$  and filtering out frequencies above  $\omega_H / \lfloor \omega_H / (\omega_H - \omega_L) \rfloor$ .

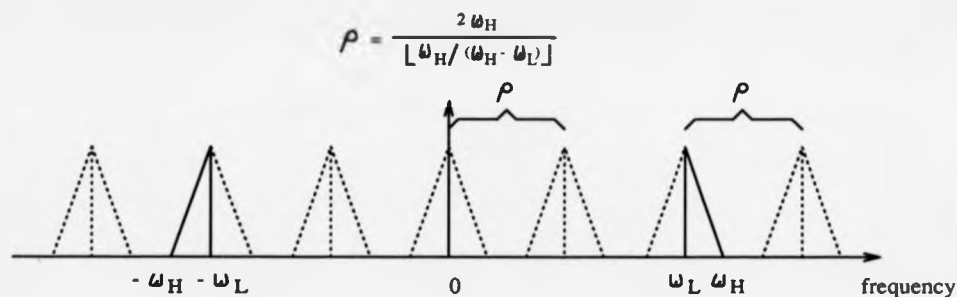


Figure 3.2: Spectrum of band-pass function

When  $\omega_H$  is an integer multiple of the bandwidth  $\omega_H - \omega_L$ , the sampling frequency is exactly  $2(\omega_H - \omega_L)$  or twice the bandwidth and is thus exactly comparable with Shannon's theorem which gives  $2\omega_H$  as the sampling frequency (Nyquist rate). When  $\omega_H$  is not a multiple of the bandwidth, the sampling frequency is  $2\omega_H / \lfloor \omega_H / (\omega_H - \omega_L) \rfloor$  and so can be thought of as about twice the bandwidth of the signal, and will be less than  $2\omega_H$ .

### 3.3 Optimal Sampling Rate Algorithm for Multi-Band-Pass Signals

#### 3.3.1 Time Domain Compression Technique

As it has been stated earlier, the more efficient the sampling process, the greater is the potential efficiency of the data processing procedure. It has been shown [Dodson, *et al.* (1985)] [Dodson, *et al.* (1988)] that larger sampling intervals can be used when the spectrum has suitable gaps: the corresponding interpolation function is the inverse Fourier transform of the characteristic function of the set outside which the Fourier transform (spectrum) of the signal vanishes.

Assuming first that the spectrum of an analogue signal is confined to a number  $n$  of intervals  $(c_j, d_j)$  (and  $(-d_j, -c_j)$ ), where  $j = 1, 2, \dots, n$  and  $c_1 < d_1 < c_2 < d_2 < \dots < c_n < d_n$ . It is required to find a number  $\rho$  such that translates of all the frequency bands by integer multiples of  $\rho$  are all disjoint. This can be done by computing the difference set of the support of the spectrum, i.e. the set of all differences of points in

$$\bigcup_{i=1}^n \{(-d_i, -c_i) \cup (c_i, d_i)\} \quad (3.5)$$

The difference set is symmetric about the origin and therefore only the positive part needs to be considered. The difference set on the positive frequency axis for the  $n$  intervals is a set of at most  $(n^2 + n + 1)$  intervals

with endpoints  $C_k, D_k$ , given by all possible differences  $(c_j \pm c_i, d_j \pm d_i)$  and  $(c_j \pm d_i, d_j \pm c_i) (i \leq j)$ , taking account of overlaps.

The sampling frequency can be calculated as follows:

$$\rho_1 = 2 \sum_{i=1}^n (d_i - c_i) \quad (3.6)$$

If  $r\rho_1$  does not land in any interval  $(C_k, D_k)$  for all non-zero integer  $r$ , i.e

$$r\rho_1 \notin (C_k, D_k) \quad 1 \leq k \leq N, r = \pm 1, 2 \quad (3.7)$$

then  $\rho_1$  is the best possible sampling rate (and  $1/\rho_1$  is the best possible sampling interval). Otherwise, suppose  $r\rho_1$  falls in  $(C_r, D_r)$  (so that  $C_r < r\rho_1 < D_r$ ) and thus define

$$\rho_2 = \frac{D_r}{r} = \rho_1 + \frac{(D_r - r\rho_1)}{r} > \rho_1 \quad (3.8)$$

and continue the process. It suffices to consider only positive  $\rho \leq 2W_{max}$ .

The process terminates after a finite number of trials. Now suppose  $\rho_{j-1}$  has been constructed successively from  $\rho_1$  and that the first interval that multiples of  $\rho_{j-1}$  fall in is  $(C_r, D_r)$ . Let  $k\rho_{j-1}$ , where  $k = 1, 2, \dots$ , fall in  $(C_r, D_r)$ ; then  $\rho_j$  is defined by

$$\rho_j = \frac{D_r}{k} = \rho_{j-1} + \frac{(D_r - k\rho_{j-1})}{k} > \rho_{j-1} \quad (3.9)$$

Since  $\rho$ 's are increasing and are, at most,  $2W_{max}$  the process terminates.



The smallest (first) value of  $\rho$  and other values are respectively the best and other possible sampling rates.

An example is shown in Figures 3.3(a)-(c): here the signal frequency lies in the intervals (300 Hz,700 Hz), (1200 Hz,1500 Hz) and (3200 Hz,3600 Hz) (Figure 3.3(a)). The corresponding difference set is in the intervals (0 Hz,400 Hz), (500 Hz,1400 Hz), (1500 Hz,3300 Hz), (3500 Hz,4300 Hz), (4400 Hz,5100 Hz) and (6400 Hz,7200 Hz). (Figure 3.3(b)). From equations 3.6-3.9, the optimal sampling rate is calculated as follows:

The possible minimum rate is

$$\rho_1 = 2 \sum_{i=1}^6 (d_i - c_i) = 2200 \text{ (Hz)} \quad (3.10)$$

if  $r = 1$  then  $r\rho_1 = 2200$  Hz falls into the difference set interval (1500 Hz,3300 Hz), so  $\rho_1$  is not the optimal sampling rate. Then from equation 3.8 the next possible sampling rate can be calculated as

$$\begin{aligned} \rho_2 &= \frac{D_1}{r} = \rho_1 + \frac{(D_1 - r\rho_1)}{r} \\ &= \frac{3300}{1} = 2200 + \frac{3300 - 1 \times 2200}{1} = 3300 \text{ (Hz)} \end{aligned} \quad (3.11)$$

When  $r = 1$ ,  $r\rho_2 = 3300$  is out of any of the difference set intervals. When  $r = 2$ ,  $r\rho_2 = 2 \times 3300 = 6600$  Hz which falls in the sixth interval of (6400 Hz,7200 Hz). Repeating the same procedure

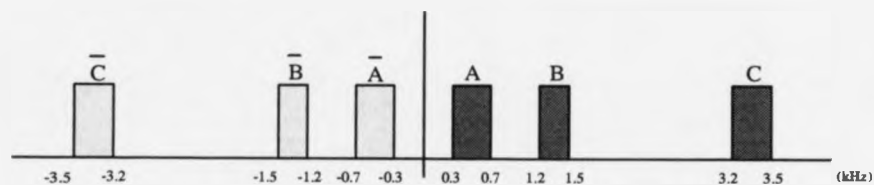
$$\begin{aligned}
 \rho_3 &= \frac{D_6}{r} = \rho_2 + \frac{(D_6 - r\rho_2)}{r} \\
 &= \frac{7200}{2} = 3300 + \frac{7200 - 2 \times 3300}{2} = 3600 \text{ (Hz)} \quad (3.12)
 \end{aligned}$$

When  $r = 1$ ,  $r\rho_3 = 3600$  Hz, again it lands in the fourth interval of (3500 Hz, 4300 Hz), so

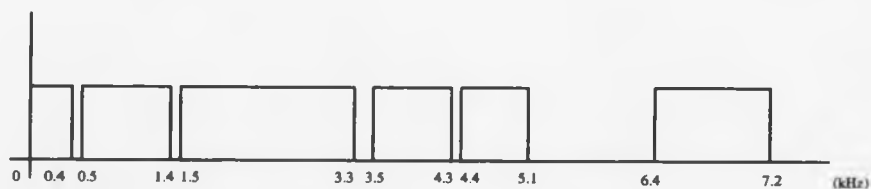
$$\begin{aligned}
 \rho_4 &= \frac{D_4}{r} = \rho_3 + \frac{(D_4 - r\rho_3)}{r} \\
 &= \frac{4300}{1} = 3600 + \frac{4300 - 1 \times 3600}{1} = 4300 \text{ (Hz)} \quad (3.13)
 \end{aligned}$$

and  $r\rho_4$  never lands in any of the difference set intervals. So the optimal sampling rate for this particular spectrum is 4300 Hz. The Nyquist sampling frequency is 7200 Hz.

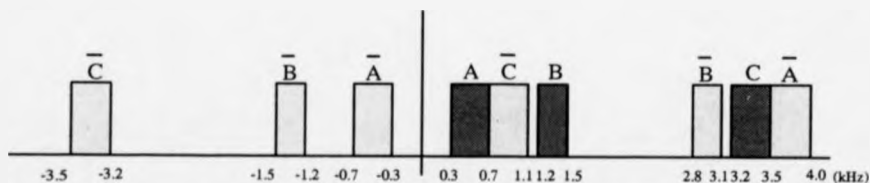
Figure 3.3(c) shows the sampled signal spectrum (only the bands between 0 and 4000 Hz are shown in detail). It can be seen that there is no overlapping of bands, even though a sub-Nyquist sampling rate was used. Since the spectrum and its translates by the sampling frequency repeat periodically along the frequency axis with period equal to the sampling frequency, some spectral components will fall into the gaps. At the receiver, these unwanted components must be removed. This requires additional information to be used to indicate to the receiver the original band positions.



(a) Multi-band-pass signal



(b) Difference sets



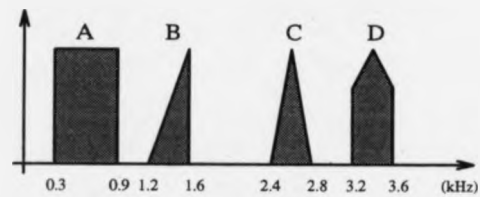
(c) Sampled signal spectrum (sampled at 4300 Hz)

Figure 3.3: Example of sub-Nyquist sampling

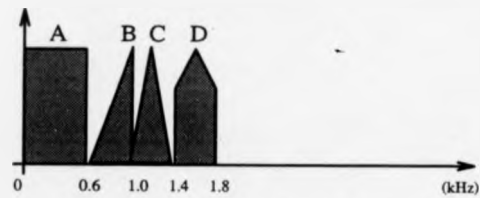
### 3.3.2 Frequency Companding Technique

Compared with the Time Domain Compression technique, Frequency Companding is a much more simple and straight forward procedure. A multi-band-pass signal can be compressed to form a low-pass signal with a bandwidth approximately equal to the total frequencies occupied by the total number of bands. This technique can be explained by an example given in Figure 3.4. This frame of signal has a spectrum which contains four bands A,B,C and D with frequency ranges 300-900, 1200-1600, 2400-2800 and 3200-3600 Hz respectively (Figure 3.4(a)). These bands can be moved down to form a compressed signal. Therefore, band A will occupy 0-600 Hz band B 600-1000 Hz, band C 1000-1400 Hz and band D will occupy 1400-1800 Hz. The compressed signal is now contained in the frequency range from 0 to 1800 Hz (Figure 3.4(b)). The bandwidth of the original signal has therefore been approximately halved. When transmitted in this compressed form the signal needs only half of the original bandwidth. The corresponding Nyquist rate, i.e. twice the compressed bandwidth, is only half of the original Nyquist rate. At the receiver, the spectrum of the received signal is expanded back to its original format to yield an estimate the original signal frequency spectrum, as shown in Figure 3.4(c).

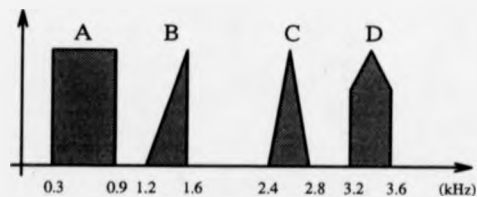
In the above frequency companding technique, additional transmitted information is also needed to inform the receiver about the position of the original bands.



(a) Original spectrum



(b) Compressed spectrum



(c) Expanded spectrum

Figure 3.4: Frequency Companding of the speech signal

### 3.3.3 Comparison of Time Domain Compression and Frequency Companding Techniques

Both the sampling reduction techniques are based on the premise that the signal has suitable gaps between its spectral elements. The main differences are as follows:

- (i) In the Time Domain Compression technique, the Nyquist sampling rate is modified. The minimum sampling rate for a frame of speech does not only depend on the maximum frequency of the signal but also depends on the number of bands, the width and location of each band. The algorithm allows the overlap between the translates of the speech bandwidth (0-4000 Hz), but by making use of the gaps it guarantees that overlap will not occur between any of these bands. Although some unwanted bands fall into the gaps, they can be removed at the receiver by the knowledge of the locations of the original bands. However, in the Frequency Companding technique, the sampling rate is still based on the Nyquist criterion. In Time Domain Compression, the minimum sampling rate is twice that of the total occupied frequency, it is likely to be higher than the minimum rate. In the Frequency Companding, all the bands are shifted down to form a baseband signal, so the sampling rate is always the minimum one.

- (ii) In Time Domain Compression, the speech bandwidth is not compressed. The original signal is sampled at the calculated optimal sampling rate, such that the band overlapping does not occur in the sampled signal (Figure 3.3). When the additional components are removed at the receiver, the reconstructed speech will be a close approximation to the original speech. In the Frequency Companding technique, the speech bandwidth is first compressed to form a baseband signal. It is the compressed signal that will be sampled and transmitted.
- (iii) From the system point of view, the Time Domain Compression requires a complicated adaptive multi-band-pass filter. A relatively simpler adaptive low-pass filter, however, is needed for the Frequency Companding system but appropriate band shifting is needed at both transmitter and receiver to compress and expand the bandwidth respectively.

The Time Domain Compression algorithm has other applications, such as speech scrambling. A speech scrambling system has been developed based on this algorithm and it is described in Chapter 5.

### 3.4 Extraction of a Multi-Band-Pass Signal from a Speech Signal

Figure 1.2 shows a typical speech waveform representing the phrase of 'HF systems are' from the sentence of 'HF systems are for users to use not for operators to operate'. The variety of structure associated with the various speech sounds is very obvious and some information about the phonetic content can be derived from the waveform plots. However, the speech waveform does not show a very important property - resonances and their time variation. The short-time spectrum of the speech signal, equivalent to the magnitude of a Fourier transform of the waveform, will be more suitable for displaying the resonances. Because the time variations of the resonances are responsible for carrying the phonetic information that results from moving the articulators, it is important to display and analyse the succession of spectra at short time intervals. Figure 1.3 shows such a short time spectrum for the speech signal in Figure 1.2 at 32 ms apart. The formants in voiced speech and their time variation are shown clearly in this spectrum. Due to the formants, the speech spectrum contains several peaks in certain frequency ranges which contain most of the energy and in the rest of frequency range the spectral components are either empty or contain very little energy. With regard to speech perception, it is the spectral peaks and not the spectral troughs which are the most significant. If the frequency components in the troughs are removed from the signal, the speech quality will not be degraded significantly.



A dynamic band extraction technique is proposed to eliminate these trivial frequency components so a multi-band-pass signal can be obtained without significantly degrading the quality. This can be described in Figure 3.5.

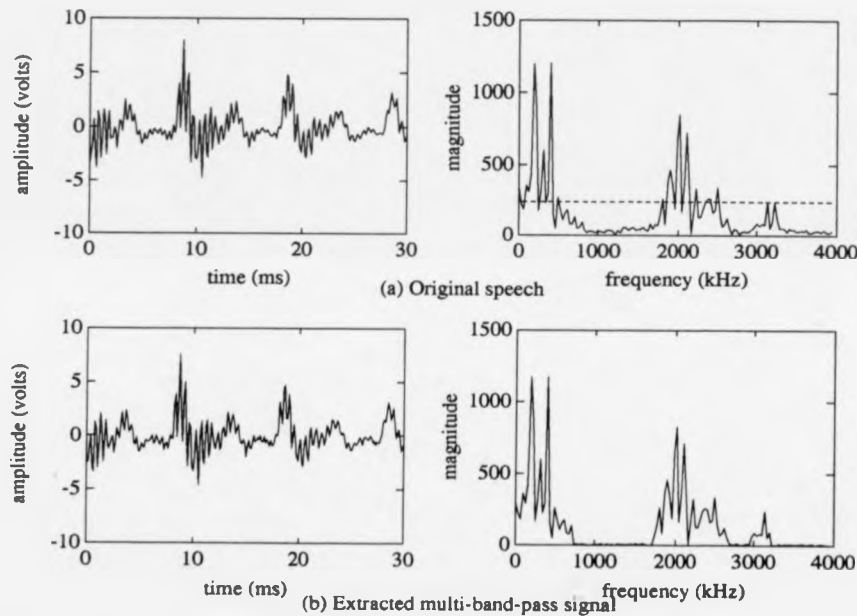


Figure 3.5: Elimination of redundancies from speech spectrum

The input speech is first sampled at Nyquist rate of 8000 Hz. The FFT is then taken for the sampled speech in a fixed period of time (in this example, it is 32 ms) to obtain the spectrum (Figure 3.5(a)). As shown in this particular frame of the speech signal, most of the energy is concentrated in the frequency ranges of 0-800 Hz, 1800-2800 Hz and 3100-3300 Hz due to the formants. A threshold can be employed to remove the trivial frequency components. The

spectrum magnitude (power) is compared with the defined threshold, and only those components whose magnitude are above the threshold are kept and those component which has magnitude lower than the threshold are deleted. By carefully choosing the threshold level, the speech quality should not be degraded significantly. As it is shown in Figure 3.5(b), the signal with redundancy removed has an almost identical waveform to the original waveform although the total frequency range occupied is only half of the original bandwidth. It must be made clear that the best reconstructed speech that can be obtained is the extracted signal. The choice of the threshold will directly affect the reconstructed speech quality and the ratio of bandwidth reduction.

### 3.4.1 Threshold Selection

Selecting an appropriate threshold level is a very important factor in deciding the final sampling rate and, hence, the reconstructed speech quality. In an adaptive-rate sampling system, the threshold could be adjustable so that a defined speech quality could be obtained at an optimal (i.e. minimum) sampling rate. The channel state, i.e. available capacity, noise characteristics etc., will also affect the threshold decision. Real-time analysis of the speech spectrum, enable the threshold to also be changed in response to the signal characteristics. Because the speech properties change from frame to frame, a fixed threshold is not practical to achieve a bandwidth reduction. Instead the threshold level has to be adjusted in each frame. As an example,

Figure 3.6 show the waveform of the word of 'city' read by an adult English male. The letters under the waveform indicate the waveform structure for the corresponding pronunciation. To analyse the spectrum for both voiced speech and unvoiced speech, two frames are taken from the sounds /i:/ and /t/ respectively. Each frame contains 256 samples equivalent to the 32 ms signal. Figure 3.7(a) and (b) show the spectrum. It can be seen that for the voiced sound /i:/ the average magnitude is much higher than that in the unvoiced sound /t/. Suppose a fixed threshold level is to be used for all the frames of this word. If the threshold is too low, such as at a scale of 50 in magnitude, almost all the frequency components in the sound /i:/ will have magnitude above this level, so no frequency component can be removed as redundancy. If the threshold is increased to 200, the frequency components in the range of [800 Hz,1700 Hz] and [3000 Hz,4000 Hz] can be eliminated to achieve a bandwidth reduction without degrading the quality significantly. However at this threshold level there is a problem for the unvoiced sound /t/. As all the frequency components are lower than the defined threshold, all the spectrum will be deleted and this results in a total loss of the information for the unvoiced speech frame. Therefore, in order to remove the redundancy but without degrading the quality, the threshold must first be decided by the characterisation of the signal.

Instead of using a fixed threshold, a dynamic thresholding is proposed. For each frame of speech spectrum, the maximum magnitude is detected. The threshold level can be defined as a percentage of the maximum value.

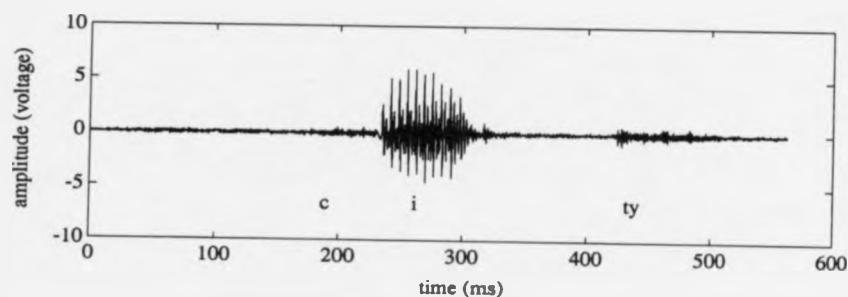


Figure 3.6: Waveform of the word of 'city'

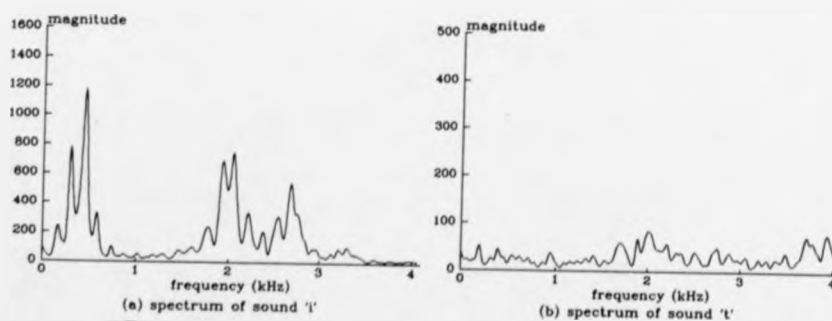


Figure 3.7: Spectrum for voiced and unvoiced sound in the word of 'city'

Generally, in a voiced speech frame, the differences between the maximum magnitude and those trivial frequency components is significant. By carefully choosing the threshold level, the redundancy can be removed and a significant bandwidth reduction can be achieved. This has been shown in Figure 3.5. In this example, the threshold is chosen as 20% of the maximum magnitude, and the total frequency range, in which the trivial frequency components are deleted is 2000 Hz, which is half of the original bandwidth. The extracted multi-band-pass signal waveform is almost identical to the original speech waveform (see Figure 3.5(a) and (b)).

Removing redundancies by eliminating trivial frequency components only makes sense in the case of voiced speech. For unvoiced speech, due to lack of formants, the spectrum is rather flat along the frequency axis (Figure 3.7(b)). If the threshold is chosen as 20%, for example, almost all the frequency components will have magnitudes above the threshold. Therefore, there will be no bandwidth reduction. If the threshold is too high, say 50% of the maximum magnitude, the threshold level will be higher than most of the magnitudes, which results in the loss of too much information. It will be explained later, the threshold level usually ranges from 10 to 30 per cent of the maximum magnitude. For most unvoiced speech segments, this threshold level will not remove too much information.

### 3.4.2 Relation Between Eliminated Spectrum and Speech

#### Quality

During the redundancy elimination process, the removable redundancies for a frame of speech depend on the speech characteristics and other factors, such as channel capacity, noise characteristics etc. In the first stage, without considering the channel condition, only speech characteristics are taken into account. The reconstructed speech quality is related to the amount of energy which remains after the thresholding process. Generally, the more energy retained in speech signals, the better the reconstructed speech quality.

There is no defined relationship between the eliminated spectrum and the quality of speech. However, a statistical relation can be obtained by practical testing. Several tests have been carried out for this purpose. The original speech was low-pass filtered with cutoff frequency of 4000 Hz. It was then sampled at the Nyquist rate of 8000 Hz. The FFT of the sampled speech was taken at a fixed frame size of 32 ms (256 samples). The speech quality was measured for different threshold levels. Two measurements were recorded: (i) intelligibility vis retained bandwidth and (ii) retained bandwidth vis threshold level.

In the intelligibility test, 250 individual words and 125 sentences were used. The content of the word and sentence lists are important, and were constructed considering phonetic balance, word length, stress position, word importance, etc [Parsons,(1987)]. These test words and sentences were read

by a male speaker and a female speaker respectively and recorded on an audio tape in a virtually noise free studio. These words and sentences were then divided into 5 groups with each group containing 50 words and 25 sentences. The intelligibility test was carried out at 5 different levels of threshold. Each group of the five lists of words and sentences was used at one of the five threshold levels, so that the intelligibility scores at different threshold levels are relatively independent of each other.

Ten listeners were asked to listen to the reconstructed speech and to write it down. The rate of correctly recorded words and sentences is defined as the intelligibility score. The average score of the ten listeners is defined as the intelligibility test result.

Figures 3.8 and 3.9 show the test results. As shown in Figure 3.9, the speech intelligibility remains almost unchanged until the total bandwidth is reduced to 2000 Hz, half of the original bandwidth. The intelligibility then slowly decays with the further reduction of bandwidth. When the bandwidth is reduced to 1000 Hz the intelligibility drops sharply. This test shows that the speech with total remaining bandwidth above 1000 Hz is highly intelligible. Figure 3.9 also shows that the intelligibility for the female voice sound is lower than that for the male at a similar bandwidth. This is due to the rich frequency components in female voice.

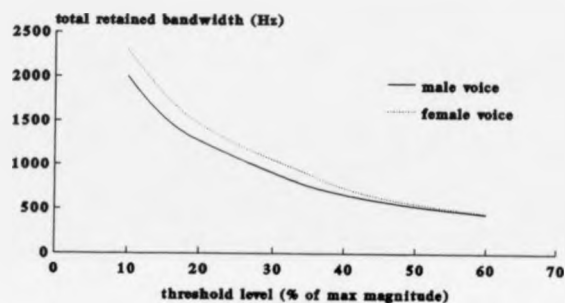


Figure 3.8: Relation between threshold level and retained bandwidth

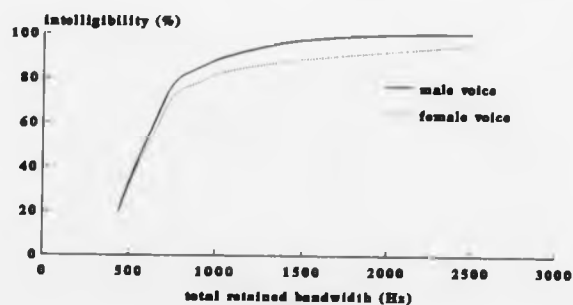


Figure 3.9: Relation between retained bandwidth and intelligibility



### 3.5 Adaptive-Rate Sampling System Design

From the above studies it can be seen that there are some redundancies which can be removed without significantly degrading speech quality. The retained speech signal, therefore, forms a multi-band-pass signal to which the sampling reduction techniques developed earlier can be applied to achieve a low rate sampling processing. The rest of this chapter describes a computer simulation to implement both Time Domain Compression and Frequency Companding techniques in speech processing. The real-time implementation using digital signal processor will be described in Chapter 6.

#### 3.5.1 System Description

The block diagrams for the Time Domain Compression and Frequency Companding systems are shown in Figure 3.10 and 3.11 respectively.

In both systems, the input speech is first sampled at the Nyquist rate of 8000 Hz to obtain a digitised signal. Then this discrete signal is divided into fixed length frame. An FFT is taken from each frame for spectrum analysis. The speech bandwidth is divided into 16 sub-bands with each sub-band occupying a bandwidth of  $4000/16 = 250$  Hz. In each frame, the maximum magnitude is detected. The threshold level varies accordingly with the maximum magnitude.

In the Time Domain Compression system, the current frame speech spectrum is fed into a spectrum analyser and a redundancy eliminator. The

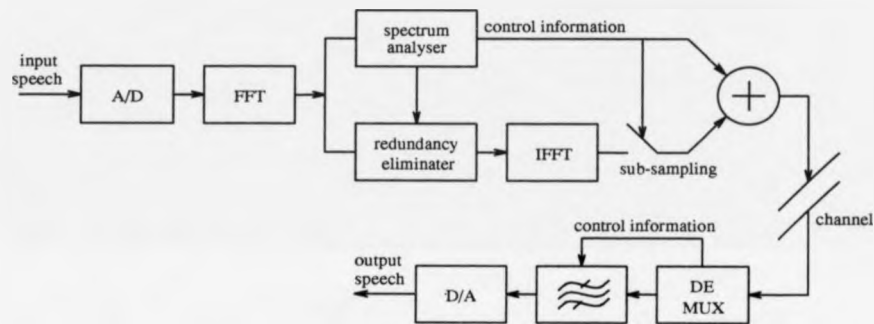


Figure 3.10: System block diagram for Time Domain Compression technique

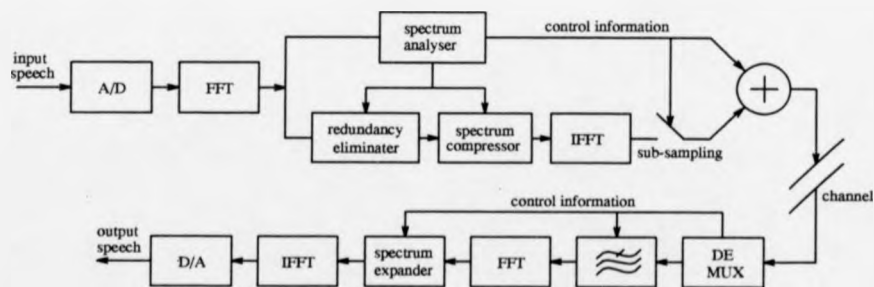


Figure 3.11: System block diagram for Frequency Companding technique

spectrum analyser detects the maximum magnitude and sets the appropriate threshold level. The speech spectrum is then compared with the defined threshold level and a decision is made that which sub-bands need to be kept and which need to be removed. The optimal sampling rate for the current frame of speech is then calculated by using equations 3.6 - 3.9. The output of the spectrum analyser is the control information which controls the redundancy eliminator to carry out the redundancy deleting. An inverse FFT is then taken from the remaining spectrum to get a time domain waveform. A sub-Nyquist sampling is taken for this signal at the calculated optimal sampling rate. The control information is also added into the speech information before it is transmitted over the channel. At the receiver, the received block of data is decoded into speech data and control data. The speech data which contains some unwanted bands in its spectrum due to the sub-Nyquist sampling is then passed through a multi-band-pass filter which is controlled by the control information. Finally, the output of the filter is converted to analogue signal to obtain a desired reconstructed speech signal.

A similar procedure is carried out in the Frequency Companding system. The redundancy is removed at the defined threshold level and the retained bands are fed into the spectrum compressor where they are spectrally compressed to form a baseband signal with bandwidth equal to the sum of total remaining sub-bandwidths. Then this baseband signal is sampled at the Nyquist rate for the compressed baseband signal. Hence the required Nyquist rate is much lower than that for the original speech signal. The re-

dundancy eliminating, spectrum compressing and sub-Nyquist sampling are controlled by the output of the spectrum analyser. This control information is also added into the speech data before transmission. At the receiver, the received data is decoded into control data and speech data. The speech signal is low-pass filtered and the FFT is taken to get the compressed spectrum. Then this spectrum is expanded back to its original frequency bands to yield approximately the original speech signal spectrum. The approximate time domain speech signal is recovered by taking an inverse FFT on this spectrum. Control information is used at the receiver for the knowledge of the original sub-band position in each frame. The low-pass filtering and spectrum expanding are both controlled by the received control information.

### 3.5.2 Simulation Results

Computer simulations have been carried out in order to verify the sampling reduction techniques. A digital signal processor (DSP32C) and PC were used for this purpose.

The analogue speech signal is first sampled at Nyquist rate (8000 Hz) and encoded with 8-bit PCM by the A/D converter in the DSP32C. The digitised speech signal is then fed into the PC for the simulation processing. The speech samples are divided into frames, and for each of these frames a sampling rate is calculated and the short-time element signal is sub-sampled at the calculated sampling rate. Because different sampling rates may be used for different frames, the number of sub-samples in each frame can be

different.

The tests were carried out at different frame size. It was found that, in general, larger frame size results in a poorer reconstructed speech quality. The reason is that, for speech signal, temporal properties such as energy, zero crossing and correlation can be assumed fixed over time intervals on the order of 10 to 32 ms [Rabiner, *et al.* (1978)]. If the frame size is too large, the speech signal may change its property from one to another during the period of one frame, for example from voice speech to unvoiced speech. When the frame is analysed in frequency domain, the unvoiced part will appear to have very low magnitude, indicating a low energy, and the voiced part will display significant magnitude, indicating high energy. When a threshold is set to eliminate the redundancies, the unvoiced speech part will be deleted. The loss of unvoiced speech signals can also happen in a small frame size case, such as 32 ms. But in this case, since the analysis frame is short, the unvoiced part of the signal in this frame will be even shorter. The loss of such a short piece of information will not be noticed by the human ear. However, in a large frame, this loss of information might be very noticeable and can severely degrade the speech quality. On the other hand, the addition of control information increases the total transmit bit rate. As the control bits are added into each frame, in order to minimise the number of added control bits, the frame size needs to be large. Making a compromise between bit rate and quality, it was found that a frame size from 16 to 32 ms (128 to 256 samples) is reasonable. The rest of the simulation results are obtained

at the frame size of 32 ms.

As an example, Figure 3.12 and 3.13 show the test results for the Time Domain Compression system and the Frequency Companding system respectively. The speech used is a short sentence of 'Is the earth really flat?' read by an adult male. The frame size is fixed at 32 ms, or 256 samples. The total sentence is 1.664 ms long and contains 13.312 samples, which means there are 52 frames in total. The threshold level is set at 20% of the maximum magnitude of each frame. Both figures show the speech waveform in different stages, i.e. input, retained signal after redundancy elimination, transmitted signal and reconstructed speech.

The results show that the average sampling rates are 4152 and 3170 Hz for the Time Domain Compression system and the Frequency Companding system respectively. Table 3.1 shows the speech classification and the corresponding optimal sampling rate for each frame of the sentence.

In general, the sampling rates for unvoiced speech are higher than those for voiced speech due to the reason described in section 3.4.1. The statistical figure for this particular sentence shows that the average sampling rate for voiced, unvoiced speech and silence are 3761, 4598 and 4992 Hz respectively for the Time Domain Compression system and 2865, 3712 and 3438 for the Frequency Companding system. The result also shows that the optimal sampling rate for the Time Domain Compression system is most likely higher than that in the Frequency Companding system. In fact only in frames 7, 8, 23 and 42, the sampling rate for both systems are the same.

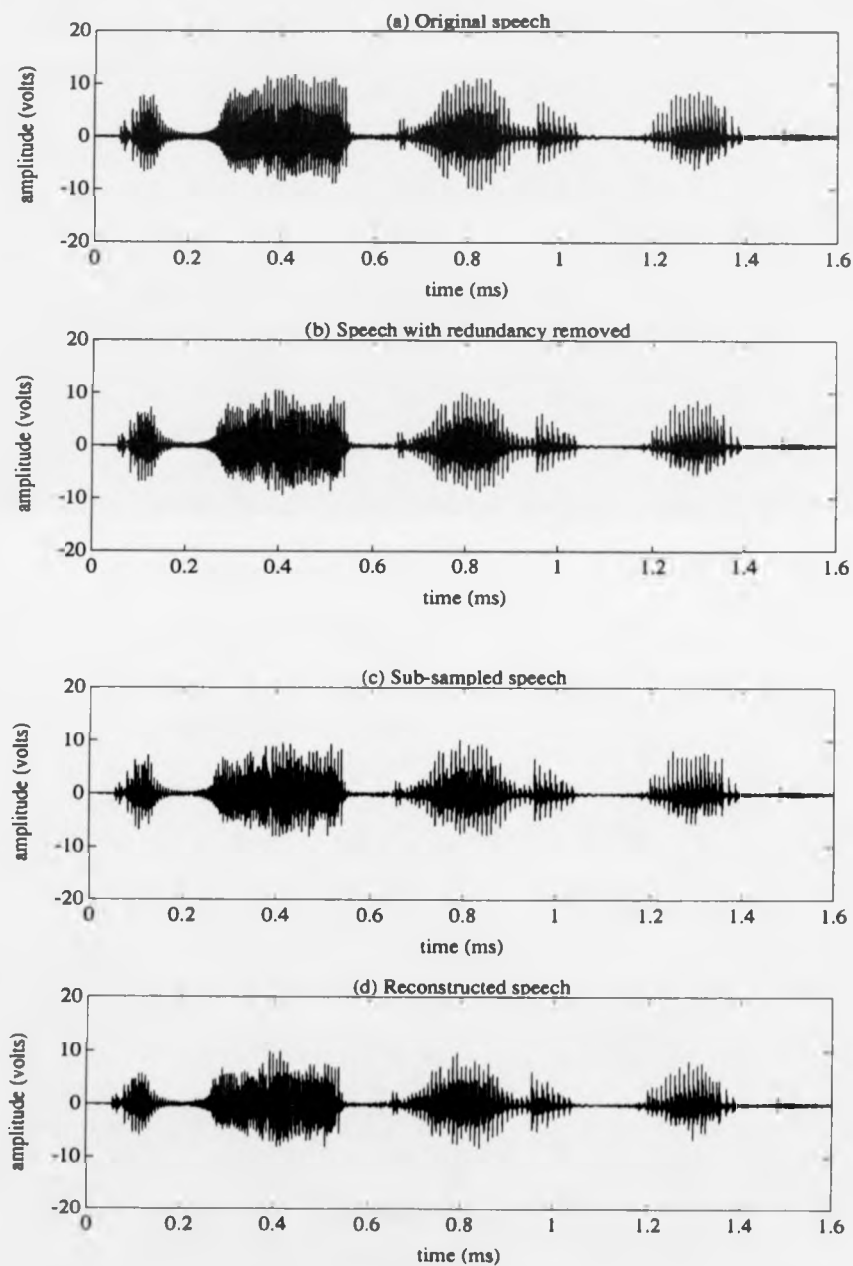


Figure 3.12: Simulation result for Time Domain Compression system

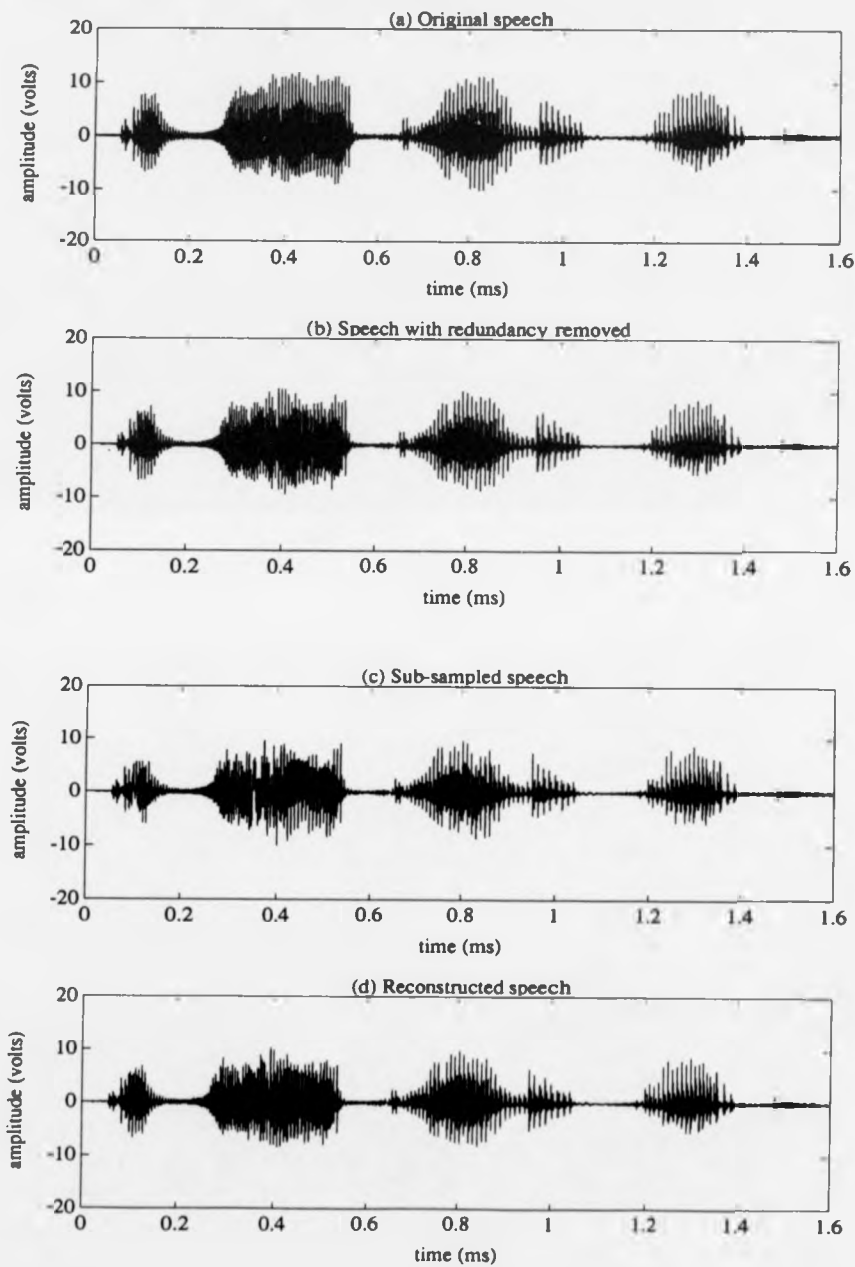


Figure 3.13: Simulation result for Frequency Companding system



In the above test, the threshold level is adjusted from frame to frame according to the speech spectrum. In a real-time system, this adjustment can be carried out based on both the speech character and the channel condition. There is always a trade off between the reconstructed speech quality and the sampling rate. By adjusting the threshold level, the sampling rate can be varied and consequently the reconstructed speech quality changed. This is very useful in real-time communication systems in which the threshold level would be adapted to the state of channel by using real-time channel evaluation (RTCE) techniques [Darnell.(1983)] [Honary, *et al.*(1988)]; the sampling procedure would continuously adjust the spectral threshold level so that the sampling rate calculated for the remaining multi-band-pass spectrum remained compatible with the available channel capacity. This will be described in detail in Chapter 6.

Frame No.	VUS status	Kept bandwidth (Hz) (after thresholding)	Sampling rate (Hz)	
			TDC	FC
1	S	768	1792	1536
2	U	768	2816	1536
3	V	768	1920	1536
4	V	1280	3712	2560
5	V	1280	3712	2560
6	U	768	1920	1536
7	U	384	768	768
8	U	512	1024	1024
9	V	1280	3072	2560
10	V	1280	3072	2560
11	V	1024	2944	2048
12	V	1024	3200	2048
13	V	1280	3968	2560
14	V	1024	3456	2048
15	V	768	2560	1536
16	V	1024	2944	2048
17	V	1024	2432	2048
18	V	1536	3328	3072
19	U	1536	4864	3072
20	U	2816	6656	5632
21	U	2560	4864	5120
22	V	768	1792	1536
23	V	512	1024	1024
24	V	768	1792	1536
25	V	1024	2816	2048
26	V	1536	4224	3072
27	V	1536	4480	3072
28	V	1536	4864	3072
29	V	1792	6144	3584
30	V	1280	3200	2560
31	V	2048	5760	4096
32	V	2048	6400	4096
33	V	1024	3840	2048
34	U	1280	4096	2560
35	S	1024	2688	2048
36	S	1536	5376	3072
37	S	2304	7552	4608
38	U	3328	8192	6656
39	V	1792	4352	3584
40	V	2304	4864	4608
41	V	2304	5120	4608
42	V	2560	5120	5120
43	V	2304	5120	4608
44	V	2816	5376	5632
45	U	4608	5120	4608
46	U	1280	3584	2560
47	U	1792	4864	3584
48	U	3840	8192	7680
49	U	2816	7424	5632
50	S	3072	7424	6144
51	S	2304	7680	4608
52	S	1024	2432	2048

Table 3.1: Speech status and the optimal sampling rate for each frame

## **Chapter 4**

# **Reconstruction of Speech from Its Frame Differences**

## 4.1 Introduction

Human voice exhibits some distinct properties, such as quasi-randomness, pseudo periodicity and silence. Among these properties, pseudo-periodicity gives a possibility of speech compression. Various techniques have been proposed for this purpose. Two examples of these techniques were described by Gieseler [Gieseler,*et al.*(1980)] and Jibbe [Jibbe,(1986)]. Basically, both of the systems take one frame of a speech signal as a reference and eliminate others which are similar to the kept one. During reconstruction, the retained frame speech simply repeats itself for desired number of times to form an approximate of the original signal. The reconstructed speech using these techniques is reported to be intelligible but lacks naturalness, sounds rather artificial, due to the fact that these techniques ignore the changes of the stress, pitch period and the speaker's emotion.

The pitch is the fundamental frequency of speech. It determines the period of the pseudo-periodical signal. Pitch variations over a sentence give shape to a sentence and indicate its structure. Stress indicates the degree of emphasis with which a word or syllable is spoken. Stressed sounds are usually louder than unstressed one. Also stress tends to raise pitch, which causes pitch change from an unstressed sound to a stressed sound. The changes of pitch, stress etc make speech sound natural and with emotion. In order to achieve a high quality speech reconstruction, these changes must be traced and, ideally, need to be preserved as much as possible. In this

chapter, the author proposes a new technique which compresses the speech bandwidth by making use of the pseudo-periodic property. Unlike other techniques, this technique is able to trace the changes of the speech and restore these changes during the speech reconstruction. The reconstructed speech is highly intelligible and sounds natural because the time-variations of the speech signal are preserved.

## 4.2 Examination of Time-Variation of Speech Signal

In order to examine the time-variation of a speech signal, the waveform which has been shown in Figure 1.2 is shown in Figure 4.1 in greater detail. This speech was sampled at 8000 Hz; the elapsed time is marked on the figure at intervals. The first thing to observe in this figure is the fact that there are virtually no abrupt boundaries between phones; nearly every sound fades gradually into the next sound. In the following discussion, the phones will be described as starting at certain points, which are identified by the letters in the figure, but it must be remembered that these are approximate locations. The phonation begins at point A. This is for the sound of /ei/ in 'H'. Here we can see the characteristic shape of the speech waveform: each cycle begins with a marked peak and is followed by a set of gradually decaying oscillations. The initial peak results from the start of the glottal pulse, and the following oscillations are the vibrations of the cavity resonances in the

mouth in response to the pulse. This sound lasts for about 125 ms, but there is an obvious change of the waveform shape. In linguistic terminology, the phone /*ei*/ is a diphthong, defined as a gliding monosyllabic speech item that starts at the articulatory position for one vowel and moves to the position for another [Rabiner, *et al.* (1978)]. The phone starts with /*e*/ in /*ei*/ then repeats for about 7 cycles with a pitch of 120 Hz. Then at point B, the transition between /*e*/ and /*i*/ starts and the pitch and formants gradually change. The pitch for the later part of /*ei*/ is about 128 Hz. Both of the pitches are reasonable figures for a male voice.

The unvoiced affricate sound /*tf*/ starts at point C and lasts for about 86 ms until point D. This is an obvious noise-like signal with much lower energy level compared with the voiced sound /*ei*/ . The sound /*e*/ of 'F' starts at point D. Compared with /*ei*/ in 'H', the formant transitions and amplitude changes of the sound /*e*/ are much smoother. From point D to E, a duration of approximately 104 ms, there are 13 cycles, so the pitch in this region is about 125 Hz. If the /*e*/ were constant, then all the cycles would have the same shape. However, the vocal cords take one or two cycles to reach a steady state; the tongue is continually in motion and the lips are moving from the open position of the beginning of /*e*/ to the closed position needed for the sound /*f*/ . These changes, which are examples of coarticulation and formant transitions, are reflected in the continually changing shapes of the 13 cycles of the sound /*e*/ .

The unvoiced fricative /*f*/ (from point E to F) is followed by another

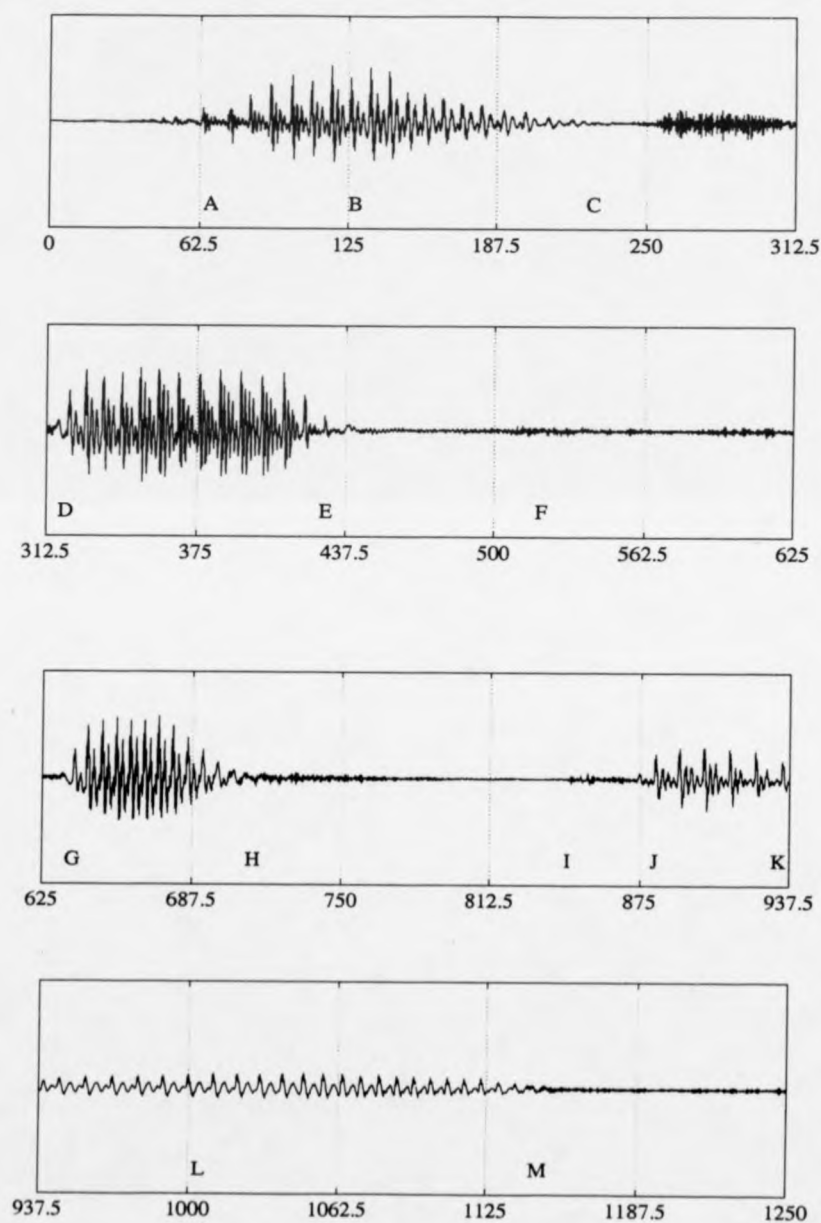


Figure 4.1: Time waveform of the phrase of 'HF systems'

fricative /s/ for the first 's' in 'systems'.

The vowel /i/ starts at point G with a pitch of 160 Hz and is followed by the sound /s/ for the second 's' in 'systems'. A short period of sound /t/ starts at I and lasts for about 28 ms. The sound /e/ and /m/ are both voiced sound. The /e/ starts at point J and then the lips move together to closure for sound /m/. At about point K, the lips are completely closed and the shape of waveform has changed from that for /e/ to that for /m/. There are five periods for sound /e/ with pitch of 98 Hz. The sound /m/ lasts relatively longer than other phones, with about 24 cycles. The effect of stress can be seen clearly here. The speaker had put stress at the later part of /m/ and this results in an increase of pitch. The /m/ starts with a pitch of 96 Hz and the period almost repeats identically for about 6 cycles. At about point L, the stress starts and the amplitude is increasing together with the pitch. The pitch is increasing gradually from 96 Hz at point L to 133 Hz at point M. Finally the sound /m/ is followed by a noise-like sound /s/ for the last 's' in the word of 'systems'.

### 4.3 Compression Technique

From the discussion in section 4.2, it can be seen that speech signal is basically a nonstationary signal. The time variations of pitch, formants and stress not only happen during the transition between two phones but also happen in the period of a single phone. Such variations are usually a slow



procedure so that the speech waveform repeats many times at the length of the pitch period and with little change between two consecutive pitch periods. If this change can be detected and extracted, one period of speech signal can be reconstructed by modifying its neighbouring period using the frame differences. This can be explained clearly in the following example.

Figure 4.2 is a piece of speech with duration of 32 ms. The pitch frequency is of 80 Hz. This signal is sampled at 8000 Hz so there are 100 samples in each pitch period. First, we consider the case of fixed analysis frame size. Suppose the frame size is fixed as 16 ms, i.e. 128 samples in each frame. The waveforms with magnitude and phase responses in the frequency domain for both frames are shown in Figure 4.3 (a),(b) and (c) respectively. The waveforms are similar but with a time shift between them (Figure 4.3(a)). This similarity also exists in the frequency magnitude which indicates the power distribution (Figure 4.3(b)). However, because of the time shift, the phase response of the signal will normally be un-correlated as shown in Figure 4.3 (c). The differential signal can be extracted in the frequency domain by comparing two frames' spectrum or, more directly, in the time domain.

The purpose of reconstructing speech from frame differences is to achieve a low sampling rate. To employ the optimal sampling rate algorithms developed in Chapter 3 to the frame differences, the differential signal should occupy a very narrow bandwidth in total. It was found that the length of the frame is a very important factor in this technique. Although the magnitudes in the two frames are similar, it is not practical to extract the differences

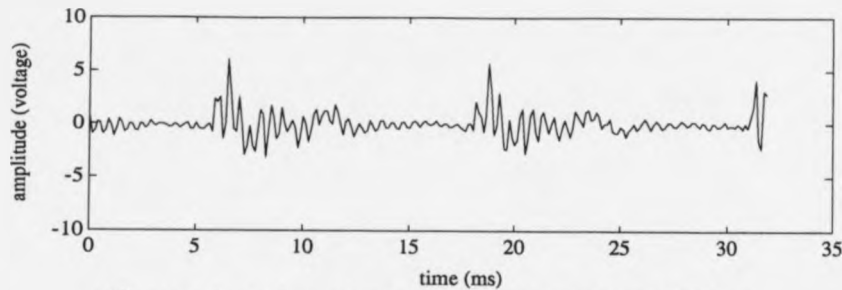


Figure 4.2: A short time segment of speech signal (32 ms)

without considering the frame size. This is because the time domain waveform in two consecutive frames may look very similar but with a possible time shift. This implies highly correlated magnitude but uncorrelated phase responses in frequency domain. A narrow-band frame differences cannot be extracted unless both the magnitude and phase in two frames are highly correlated. This can be seen more clearly if the differences are taken directly in the time domain. Figure 4.4(a) is the differential signal extracted from the corresponding samples from the two frames. In fact, the extracted differential signal looks like a normal speech signal with a typical speech spectrum 4.4(b). A threshold can be employed to remove some redundancies and sampling rate reduction can be achieved. However this is basically the same process which was described in Chapter 3, so the ratio of sampling rate reduction will be the same as was achieved in Chapter 3.

This problem can be solved, if the frame size is chosen adaptively according to the speech signal pitch period. The pitch for the signal in Figure 4.2 is 80 Hz (100 samples). If the frame size is set as the same length as

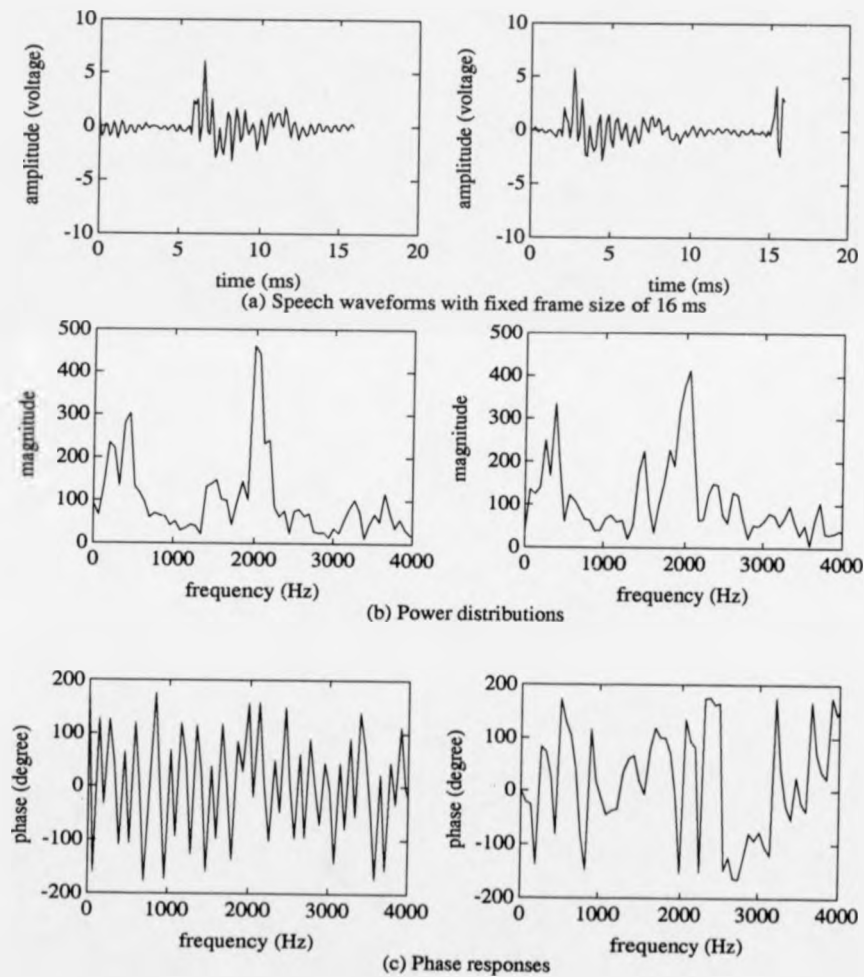


Figure 4.3: Speech waveform and frequency responses in a fixed length frame (32 ms)

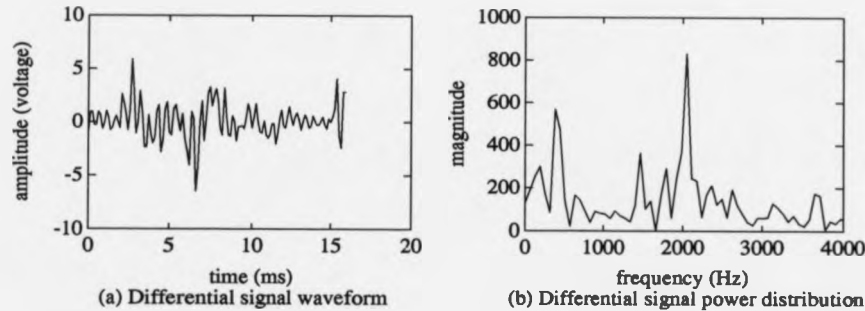


Figure 4.4: Differential signal in fixed length frame

the pitch period, the two frame signals will have similarities in both time domain waveforms and frequency responses; magnitude and phase, as shown in Figure 4.5(a), (b) and (c). The differences from the two frames can be extracted by only comparing their magnitudes.

As shown in Figure 4.6(a), the significant differences, the peak, concentrate in a small frequency range from 1650 to 2250 Hz. In other ranges, the differences are very small. It is the significant differences that make the major contribution to the changes of waveform structures between the two periods of speech signal, so if only the peak is preserved and used later on in reconstruction of the signal, the changes will be followed almost entirely. As an example, the Frequency Companding technique developed in Chapter 3 is used to compress the bandwidth. The kept band is moved down to form a baseband signal with a bandwidth of 600 Hz (Figure 4.6(b)). The inverse FFT is performed to obtain the compressed signal waveform (Figure 4.6(c)).

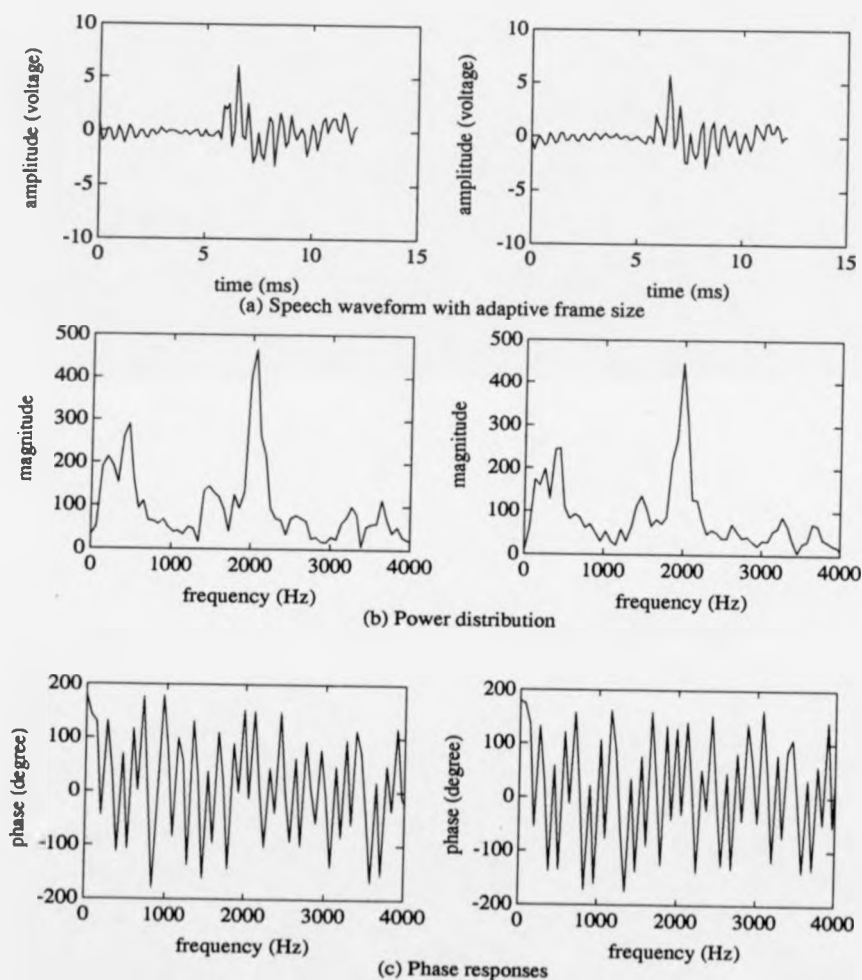


Figure 4.5: Speech waveform and frequency response in two consecutive periodic-frames

The sampling rate needed for the compressed signal is 1200 Hz. The second frame of signal can be reconstructed by simply adding the frame differences to the first frame signal (Figure 4.6(d)).

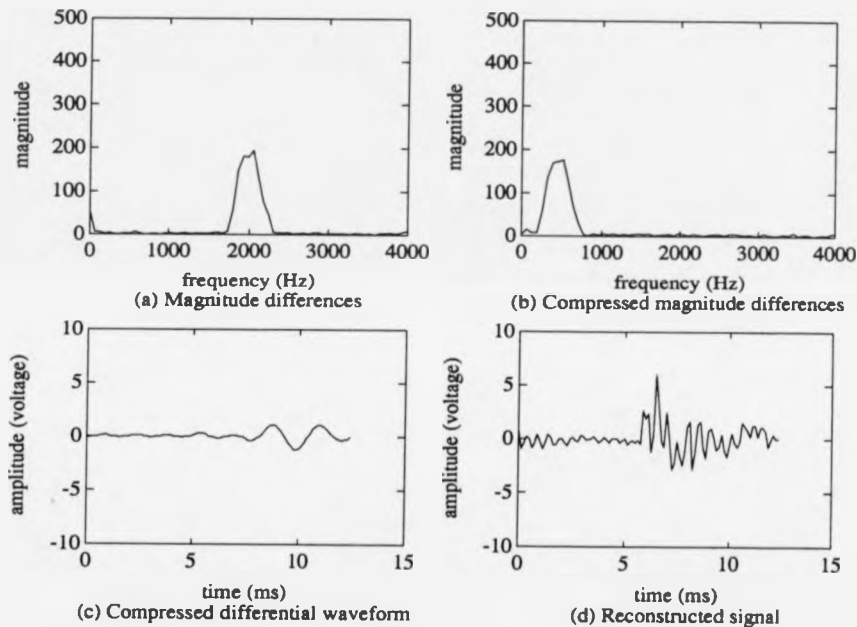


Figure 4.6: Differential signal and reconstructed signal

## 4.4 System Description

A general block diagram of the compression system is shown in Figure 4.7. The functional details of each block in this figure will be illustrated in the subsequent section of this chapter.

Referring to Figure 4.7, the speech is first sampled at the Nyquist rate

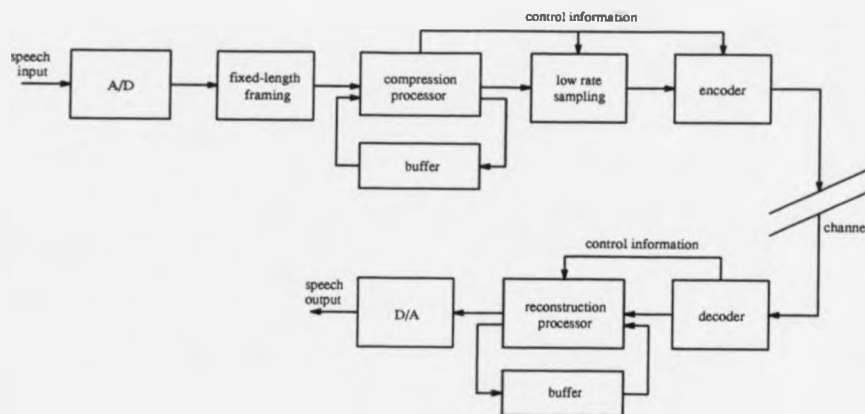


Figure 4.7: System block diagram

of 8000 Hz and encoded with 8-bit PCM. The sampled speech signal is then framed with a fixed length of 32 ms, or 256 samples. It is then determined whether each frame is voiced speech, unvoiced speech or silence. If it is voiced speech, its pitch period is detected. Each frame of speech signal is also fed into the compression processor which treats the incoming frames differently according to their VUS status. The compressed speech information bits together with some necessary control bits are sent to the receiver or stored in a long term medium for future reconstruction.

The reconstruction processor is basically an inverse processing of the compression. From the knowledge of the VUS status of the incoming signal, the processor reconstructs the speech signal by using an appropriate method. Finally the reconstructed speech is sent to the destination which could be a telephone or an amplifier-speaker unit.

A speech reconstruction process is also carried out at the transmitter, and this processing is basically the same as that at the receiver except that there is no need to compress the frame differences here. The reconstructed signal is also kept in a buffer and the next comparison should be made between the reconstructed signal in the buffer and current period, and the differences be extracted from them. This will prevent an accumulated error from adding into the reconstructed speech.

#### 4.4.1 Voiced/Unvoiced/Silence and Pitch Detector

The need for deciding whether a given segment of speech waveform should be classified as voiced, unvoiced or silence and determining the pitch period for a voiced speech has been studied by many researchers [Dubnowsk,*et al.*(1976)] [Schafe,*et al.*(1970)] [Markel,(1972)] [Rosenber,*et al.*(1975)] [Ross,*et al.*(1974)]. Pitch determination, i.e. detection and measurement of the fundamental frequency  $F_0$  of voice in natural human speech, is one of the most important subjects in speech processing. In speech coding, for instance, the quality of the vocoder speech deteriorates rapidly as a function of imprecise pitch estimates. Accurate and reliable measurement of pitch period of a speech signal from the acoustic waveform is often a difficult task for several reasons. One reason is that the glottal excitation waveform is not a perfect train of period pulses. The waveform varies both in period and in the detailed structure. A second difficulty in measuring pitch period is the interaction between the vocal tract and the glottal excitation. In some instances the formants



of the vocal tract can alter significantly the structure of the glottal waveform so that the actual pitch period is difficult to detect. A third difficulty in pitch detection is distinguishing between unvoiced speech and low-level energy voiced speech. In many cases transitions between unvoiced speech segments and low-level energy voiced speech segments are very subtle and thus are extremely hard to pinpoint.

Because of the importance of such classification and determination, a wide variety of algorithms for speech classification and pitch detection have been proposed in the speech processing literature and are reviewed in Chapter 2. Due to the difficulties in the VUS classification and pitch determination, none of these algorithms performs significantly better than the others in all circumstances. The choice of the pitch detection algorithm depends on the requirement of the user. It was found that the SIFT algorithm performs quite well for both male speakers and female speakers [Rabiner, *et al.*(1976)]. In the speech compression system using frame differences, SIFT is employed to discriminate speech segments as voiced or unvoiced speech and extract the pitch period for voiced speech.

A block diagram of the SIFT analysis system is shown in Figure 4.8. The speech waveform  $s(t)$  is first pre-filtered by a low-pass filter with a cutoff at 800 Hz since the spectrum of a voiced sample will always have a maximum peak in the range (0,1000) Hz with the largest peak outside the range, generally 5-10 dB below the first peak. A cutoff at 800 Hz is a reasonable choice for including most of the low-frequency range while providing sufficient at-

tenuation at 1000 Hz. After sampling the filter output at a 2000 Hz rate by a decimation process (i.e. 3 out of 4 samples are dropped at the output of the lowpass filter), the output,  $x(n)$ , is then analysed using the autocorrelation method. It is found that, a fourth order filter is sufficient to model the signal spectrum in the frequency range 0-1000 Hz because there will generally be only 1 – 2 formants in this range. The signal  $x(n)$  is then inverse filtered to give  $y(n)$ , a signal with an approximately flat spectrum. The short-time autocorrelation of the inverse filtered signal is computed and the largest peak in the appropriate range is chosen as the pitch period. To obtain additional resolution in the value of the pitch period, the autocorrelation function is interpolated in the range of the maximum value. An unvoiced classification is chosen when the level of the autocorrelation peak falls below a given threshold.

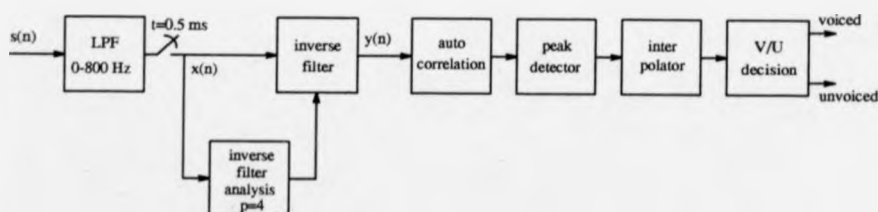


Figure 4.8: Block diagram of SIFT algorithm

Figure 4.9 shows some typical waveforms obtained at several point in the analysis. Figure 4.9(a) shows a frame of input speech being analysed. Figure 4.9(b) indicates the input speech spectrum and the reciprocal of the spectrum of the filter. For this example there appears to be a signal formant in the

range of 250 Hz. Figure 4.9(c) shows the spectrum of the signal at the output of the inverse filter, whereas Figure 4.9(d) indicates the time waveform at the output of the inverse filter. Finally, Figure 4.9(e) shows the normalised autocorrelation of the signal at the output of the inverse filter. A pitch period of about 8 ms is clearly presented by the peak value.

Similar to many other VUS classification techniques, SIFT can only discriminate voiced speech and unvoiced speech. If a segment of a signal is silence, it will be classified into either voiced speech or unvoiced speech. In practical implementation, a preliminary silence detector is employed to detect the silence in speech signal and the decision is made based on the total of energy contained in a frame of signal. Generally, voiced speech contains much higher energy than that in unvoiced and silence, and the total energy contained in unvoiced speech and in silence is also quite distinguishable. An appropriate threshold can be employed for the purpose of silence detection and any frame of speech signal which has a total energy below this threshold will be classified as silence. Table 4.1 shows the test results of SIFT technique. About 138.6 seconds speech is used for this test. The speech contains 40 sentences read at a normal speed without any noticeable silence (pause). The speech is VUS classified by the SIFT system and the results are compared with the decision made by manual analysis. As can be seen from Table 4.1, the detection error is as low as 2.42%, and most of the error occurred during the transition from one class of speech to another. Table 4.1 also indicates the proportion of voiced, unvoiced and silence in speech signal

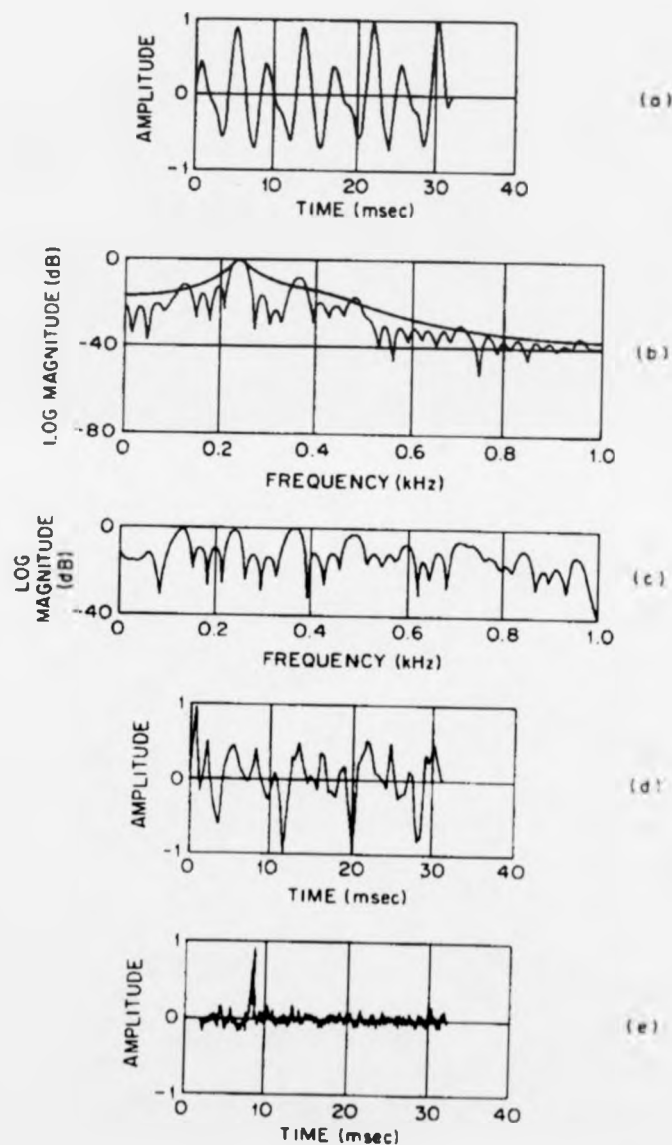


Figure 4.9: Typical signals from the SIFT algorithm ([Markel:72])

as 76.7%, 13.55%, and 9.75% respectively.

actual class identified as	voiced	unvoiced	silence
voiced	3269	43	0
unvoiced	51	538	5
silence	0	6	417
total	3320	587	422

Table 4.1: Test results of SIFT algorithm

#### 4.4.2 Compression Processor

The flow chart of the compression processor is shown in Figure 4.10. The sampled speech signal is fed into the processor in fixed length frames. Each frame is classified into voiced, unvoiced or silence by a VUS detector. Flags are used to indicate the status of each frame of signal, 0 for voiced, 1 for unvoiced and 2 for silence. Different class speech will be processed in different manner to achieve a low rate transmission or minimize the data for storage.

If a frame of signal is detected as silence, the potential of compressing the speech signal is apparent. In fact, there is no need to transmit such a piece of silence to the receiver so the whole frame data can be simply removed from the speech and only some control information sent over the communication channel. Such elimination of silence will delete all the redundancy caused by a long pause between two sentences or two words.

In the case of unvoiced speech signal, the low rate transmission of speech

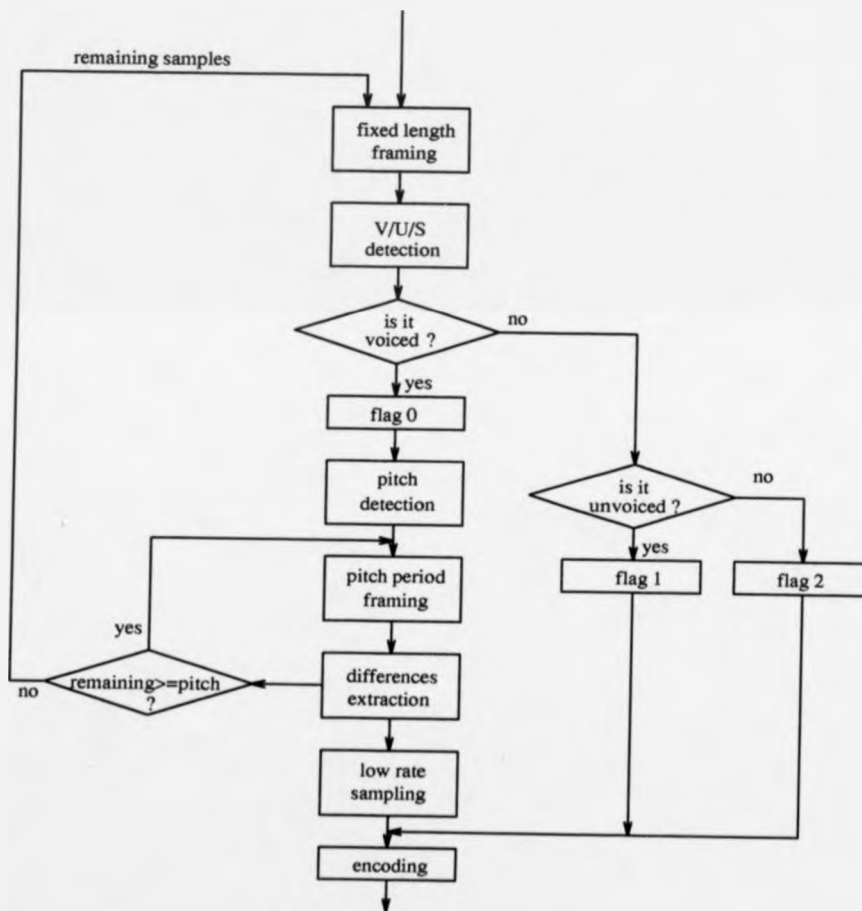


Figure 4.10: Flow chart of compression process

or minimization of stored data can be achieved by employing low bit rate speech coding techniques which were described in Chapter 2. For noise-like unvoiced speech signal, keeping its zero crossing is more important than accurately reconstructing its waveform shape. If a low bit rate speech coding technique can maintain the zero crossing information, the reconstructed signal will not show a significant loss of information. For example, the adaptive 2-bit PCM described in Chapter 2 [Turner,(1976)] is good enough to keep the original zero crossing information and roughly follow the shape of the waveform so that the reconstructed signal will not be degraded significantly.

The voiced speech is the main consideration of this system. Once a frame of speech is classified as voiced speech, its pitch period is detected by the same algorithm of VUS detection. This frame of signal is then divided into smaller frames which have a length equal to the detected pitch period. To distinguish the fixed length frame and the frame with a length of pitch period, the former is called fixed-length-frame whereas the latter is called periodic-frame. The number of the periodic-frame which can be taken from the fixed-length-frame depends on the pitch period detected. After each periodic-frame has been taken, the compression system will count the remaining samples in the fixed-length-frame. If the number of remaining samples is more than the number of samples in the periodic-frame, another periodic-frame is taken, otherwise the remaining samples will be kept and analysed together with the incoming samples in a new fixed-length-frame.

Each of the periodic-frames is compared with the previously reconstructed

periodic-frame in frequency magnitude. The significant differences are kept to generate a time domain differential signal whilst those trivial differential components are removed from the spectrum as redundancies. As it was described earlier in this chapter, in most cases the significant differences will only occupy a very small frequency range in total, so that the optimal sampling rate algorithms which were developed in Chapter 3 can be employed to achieve a very low sampling rate. Then the extracted differential signal is compressed and transmitted over a communication channel or stored in a long term medium for future reconstruction. In the compression processor, a reconstruction is also carried out. The purpose of this reconstruction is to prevent an accumulated error from adding into the reconstructed signal. The extracted differential signal is used to modify the previously reconstructed periodic-frame signal which is kept in a buffer. The result is a new periodic-frame signal which is an approximation of the current periodic-frame. Then this newly reconstructed periodic-frame replaces the previous periodic-frame in the buffer for the following differential extraction.

#### **4.4.3 Reconstruction Processor**

As is shown in the flow chart of the reconstruction processor (Figure 4.11), the first task of the reconstruction is to decode whether the incoming signal is voiced speech, unvoiced speech or silence. This discrimination is made under the indication of the control information which resides at the beginning of each frame. The incoming signal is then passed into one of the reconstruction



processors according to its VUS status.

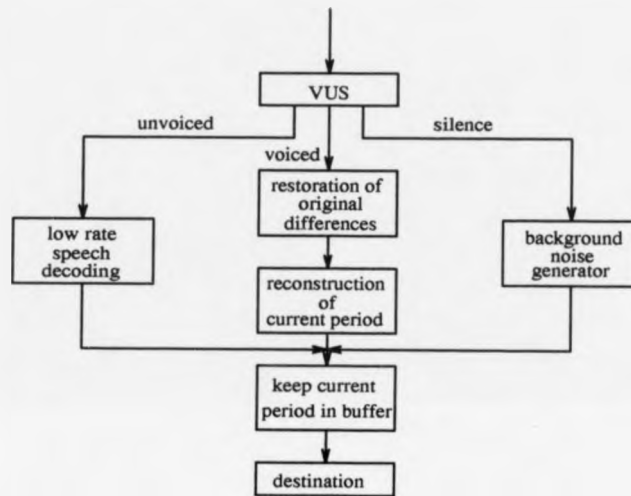


Figure 4.11: Flow chart of reconstruction process

If the incoming periodic-frame is silence, the reconstruction is simply a procedure of inserting a piece of silence (adding zero) into the signal. The length of the added silence is the size of the fixed-length-frame. In order to make the reconstructed speech sound more natural, a background noise can be generated and added into the speech as silence. The level of the noise can be adjusted manually so that the best sound effect can be achieved.

If the current periodic-frame is detected as unvoiced speech, an appropriate decoder decodes the speech data and restore the samples of the speech. The restored speech samples are then converted into an analogue signal and sent to its destination. The low bit rate transmission or storage for unvoiced speech is accomplished by using low bit rate speech coding techniques instead

of low sampling techniques.

When the reconstruction processor detects that the incoming signal is the information for a voiced speech, the bit stream for the speech data is decoded to obtain the compressed version of differential signal. By using an appropriate decompression method, the frame differences from current periodic-frame and previous periodic-frame can be restored. The differential signal is then converted into its frequency expression and its magnitude components are added to the corresponding components in the previous periodic-frame, which generates the magnitude components for the current periodic-frame. The phase responses in the previous periodic-frame are replaced by the corresponding phase in the differential signal. As the differential signal only contains the information in the frequency ranges in which the significant differences occur, the phase replacement only takes place in those ranges, whereas the phase in remaining frequency ranges will be left unchanged. The reconstructed periodic-frame waveform is obtained by taking an inverse FFT from its frequency responses of magnitude and phase. This newly reconstructed periodic-frame speech signal is stored in the buffer for the next reconstruction. Finally the speech signal is converted to an analogue signal and sent to its destination.

## 4.5 Simulation Results

A sampling rate reduction can be achieved if the technique of reconstruction of speech from its frame differences is combined with the sampling rate reduction techniques developed in Chapter 3. The results to be explained in this section are based on computer simulation. The simulation program is written in 'C' under Unix on a SUN system. The A/D and D/A conversions were accomplished by TMS digital signal processor residing on an IBM PC.

The input speech signal is firstly sampled at its Nyquist rate of 8000 Hz. The speech samples are then processed in a fixed-length-frame. The SIFT technique requires that the analysis frame must contain at least two complete pitch periods [Markel,(1972)]. Normally, the pitch ranges from 80 Hz to 160 Hz for a male speaker and from 160 Hz to 400 Hz for a female speaker. The pitch frequency is the reciprocal of pitch period, so the pitch period is from 6 milliseconds to 12 milliseconds for male and from 2 milliseconds to 6 milliseconds for female. Thus the analysis frame size is chosen as 32 millisecond, i.e. 256 samples. The fixed-length frame is classified as voiced, unvoiced or silence depending on its characters. A silence frame signal is eliminated and only control information is transmitted or stored for reconstruction. The control information for silence only indicates the class of the signal, it needs 2 bits for each frame, hence the bit rate for silence is 62.5 bit/s.

An unvoiced speech frame is encoded by using a low bit rate speech coding technique. In this simulation, a 2-bit adaptive PCM [Jayant,(1974)] is used

to encode the unvoiced speech. Again, a bit rate of 62.5 bit/s is needed to indicate the speech status. The bit rate for speech signal is  $8000 \times 2 = 16000$  bit/s, so the total bit rate for unvoiced speech is  $16000 + 62.5 = 16062.5$  bit/s. Figure 4.12 shows a piece of unvoiced speech in its original waveform and the reconstructed waveform. It shows clearly that the high frequencies were preserved perfectly.

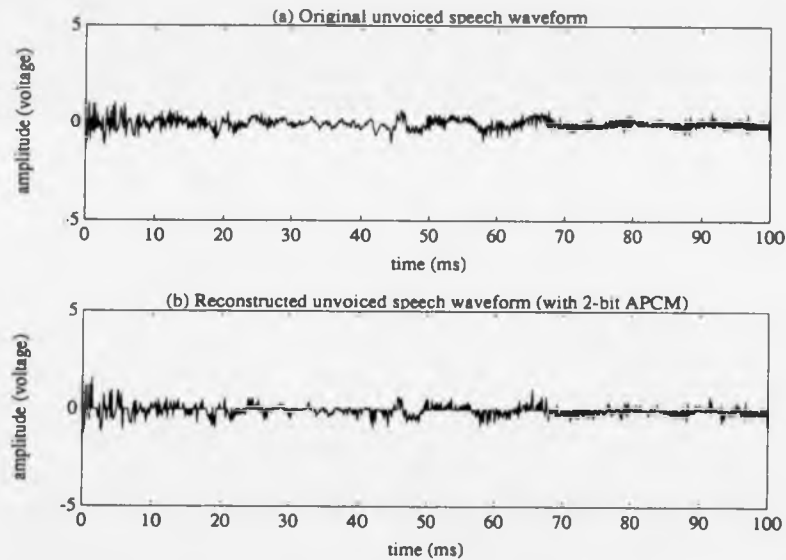


Figure 4.12: Reconstructed unvoiced speech

For voiced speech frame, the pitch period is extracted by using the same SIFT processor which was used to classify the speech status. The speech is then analysed in the frequency domain on the periodic-frame bases. The significant frame differences are extracted from two consecutive periodic-frames. The Frequency Companding technique is applied to compress the

differences and the result is converted into a time domain waveform. Because the compressed differential signal usually has very narrow bandwidth, it can be sampled at a very low rate. The simulation result shows that an intelligible speech can be obtained at an average bandwidth of 750 Hz for the compressed differential signal, hence the average sampling rate can be as low as 1500 Hz. To illustrate the detailed structure of the waveform, Figure 4.13 shows a short time segment speech signal in different stages of the compression process. The pitch periods start at the time of 40 ms and repeat for about 22 times. The waveform structure changes gradually from one frame to another. The significant differences are taken from two consecutive periodic-frames and the time domain waveform of the compressed version of the differences is shown in Figure 4.13(b). This differential signal is sampled at an average sampling rate of 1500 Hz and transmitted. Figure 4.13(c) shows the reconstructed speech waveform. It can be seen that the waveform structure appears to be very similar to the original one. The bit rate for voiced speech includes that for speech information and control information. The control information consists of 62.5 b/s for speech status and 500 b/s for indicating the position of bands. The bit rate for the speech signal itself depends on the average sampling rate, here it is 1500 Hz; if 8 bit PCM is used for encoding the samples, the average bit rate for the speech signal is  $1500 \times 8 = 12000$  b/s, and the total bit rate for voiced speech is  $12000 + 500 + 62.5 = 12562.5$  b/s.

The overall bit rate depends on the proportion of the classes present in a speech signal. From the experimental results shown in Table 4.1, we can see

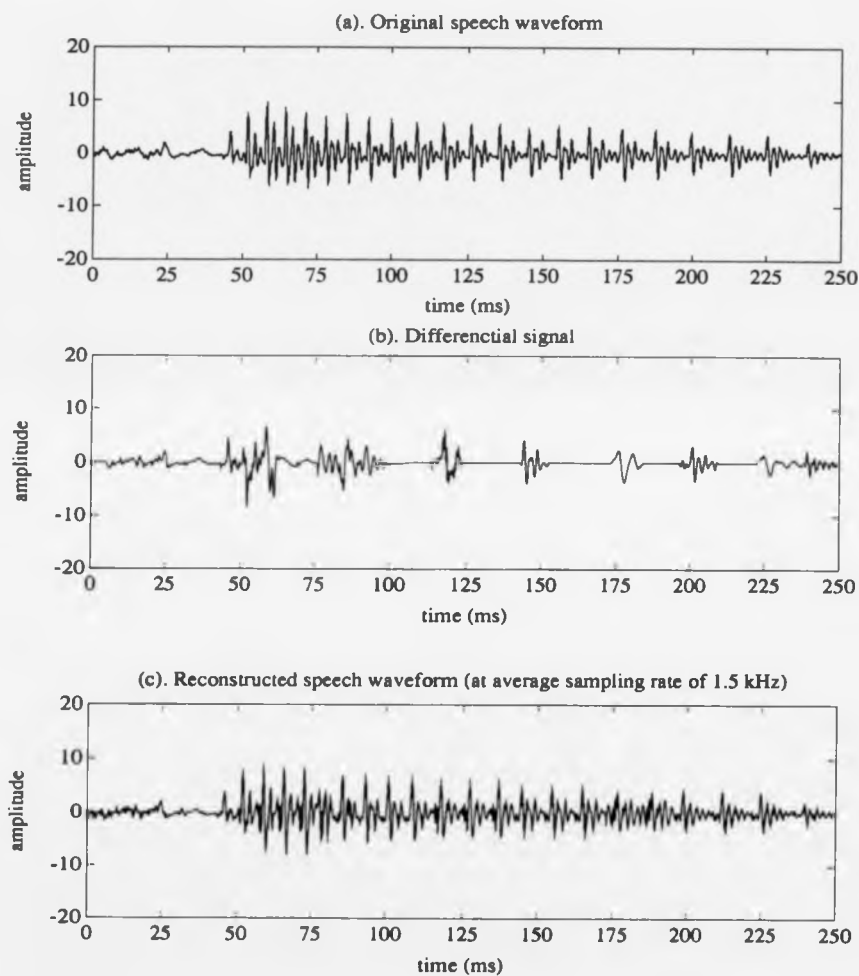


Figure 4.13: Reconstructed voiced speech

that the proportion of voiced speech, unvoiced speech and silence is 76.7%, 13.55% and 9.75% respectively. The overall bit rate can be calculated as follows:

$$76.7\% \times 13000 + 13.55\% \times 16062.5 + 9.75\% \times 62.5 = 12150 \text{ b/s} \quad (4.1)$$

The voiced speech constitutes the major part of the speech signal. In the above example, the bit rate reduction is achieved by reducing the sampling rate alone. There is considerable potential of reducing the bit rate further by combining the low sampling rate processing with a low bit rate technique.

Figure 4.14 shows a simulation result of a complete sentence of 'HF systems are for users to use'.

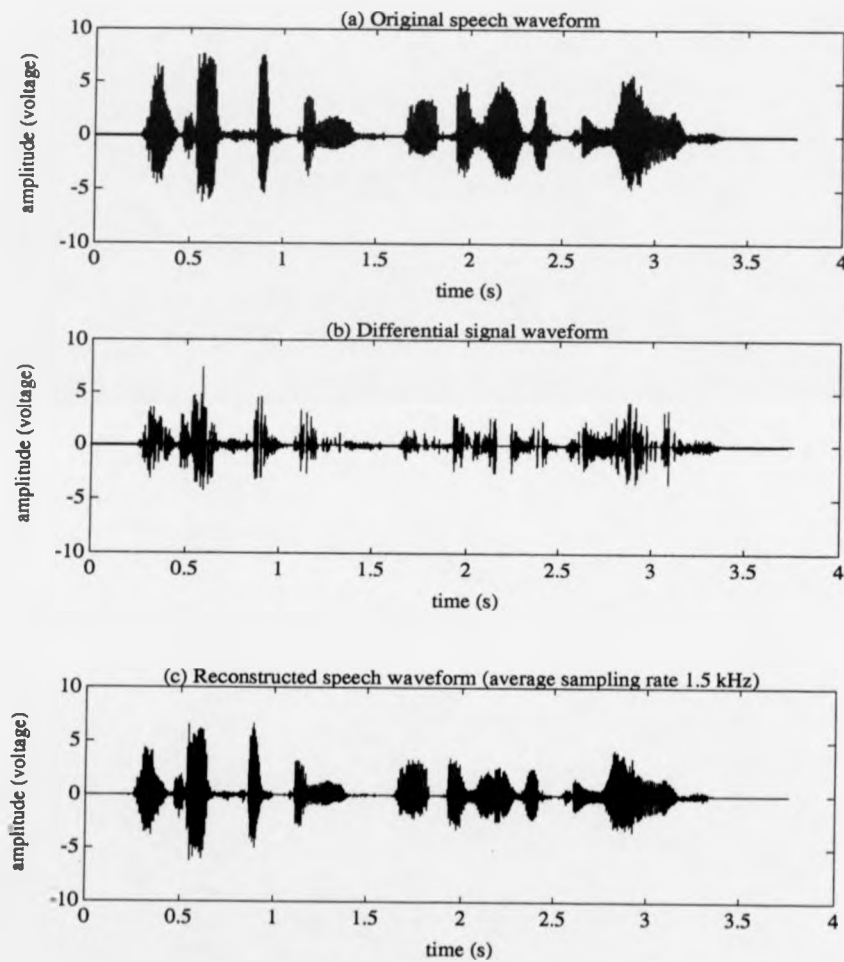


Figure 4.14: Simulation result of reconstruction of speech from frame differences



## **Chapter 5**

# **Speech Scrambling Employing Adaptive-Rate Sampling**

## 5.1 Introduction

When a message signal is transmitted over a communication channel, the signal can be accessible not only to the intended receiver but also to any unauthorised party who wants to recover the transmitted message. It is often desirable to achieve privacy in a communication system that uses such an 'open' channel as the transmission path. Such privacy can be achieved by the process of speech scrambling.

Prior to transmission, a scrambling process is performed on the original speech, which has the effect of disguising or masking its content. The intended receiver can effectively recover the original message by performing descrambling which reconstructs the original message or a reasonable approximation to that message.

In order to prevent any unauthorised reconstruction of a transmitted message, a large set of possible transformations is made available, from which one is selected for a limited period time. Corresponding to each transformation, a particular descrambling transformation is needed at the receiver. The particular transformation agreed upon by the transmitter and the receiver is identified by a parameter called the 'key'. The number of possible transformations is called 'key size'.

In speech scrambling, the degree of security is measured by two factors: the 'key size' and the 'residual intelligibility'. The residual intelligibility is a measure of the remaining intelligible message in the scrambled speech.

Generally, a large 'key size' and a low residual intelligibility means a high degree of security.

Speech scrambling may be classified under two major headings [Del Re, *et al.*(1989)]:

- (1) **Analog Scrambling:** In analog scrambling, the only real analog operation is signal transmission since processing is carried out digitally [Jayant.(1982)]. Incoming speech signals are digitised, processed using a special algorithm, converted to analog form, and transmitted to a receiver, where they are digitised again, inversely processed, and reconverted to analog form for reconstruction. The analog speech scrambling has less bandwidth expansion and are applicable over the existing telephone channels with standard telephone bandwidth at acceptable speech quality [Lee,*et al*(1984)]. However, it is far less secure than digital speech encryption, due to the limitation of time and frequency domain scrambling approach [Jayant, *et al.*(1981)] [Jayant, *et al.*(1983)] [Baschlin,(1977)] [Brunner,(1980)] [French,(1973)]. The residual intelligibility is in the range of 20% – 70% [Jayant.(1982)].

- (2) **Digital Encryption:** In digital encryption, the signal is digitised and compressed to reduce the bit rate [Jayant.(1982)]. The cipher modifies the sequence bit series by means of block or stream ciphering [Branstad.(1978)]. The modified sequence is then transmitted via digital modulation. The digital speech encryption usu-

ally provides high-level security (low residual intelligibility). The residual intelligibility in encrypted speech can be as low as zero [Orceyre, *et al.*(1978)]. But the digital speech encryption suffer a serious drawback of bandwidth expansion [Del Re, *et al.*(1989)], while low bit rate speech coding technique is required to compress the speech before the encryption taking place. This implies a relatively poor recovered speech quality.

In this chapter, a brief review is given on the issue of the existing analog scramblers for speech privacy. A new scrambling algorithm is then proposed to achieve a high degree of security analog speech scrambling. In this new speech scrambling system, speech is spectrally scrambled at the transmitter to produce a defined multiple band pass structure. The reconstruction 'key' comprises a knowledge of the spectral frequency bands of the scrambled elements together with a unique sampling rate which will allow correct recombination of the elements. To prevent unauthorised reconstruction, the bandpass structure and sampling rate can be changed irregularly according to an appropriate key sequence. Both single user and multi-user systems are designed. Also the sampling rate tolerance is calculated.

## 5.2 A Brief Review of Speech Scrambling Techniques

Analog speech scrambling can be classified as one dimensional or two dimensional [Jayant, *et al.*(1981)]. One dimensional algorithms are those that manipulate a signal in either time domain or frequency domain, while two dimensional algorithms are combination of two or more one dimensional algorithms.

### 5.2.1 One Dimensional Speech Scrambling

The analog scrambling techniques started from the basis of speech signal manipulations in the frequency domain [French,(1973)]. The techniques provide a satisfactory degree of transmission robustness in the context of the real channel operation, but with a high residual intelligibility [Jayant.(1982)]. The advancement in digital signal processing techniques has made the time domain speech scrambling become more practical. The possibility of a variation of temporal manipulation in these techniques, the destruction of the rhythm of speech and the possibility of mixing up segments of active speech and silence, make learning of time-manipulated speech less tractable than the learning of frequency-manipulated speech [Jayant, *et al.*(1981)]. However, realisation of the low levels of residual intelligibility possible in these techniques demands the use of extremely large values of communication delay [Jayant, *et al.*(1981)].

### Frequency Domain Scrambling

**Frequency Inversion** [Nelson,(1976)]: The frequency inversion is illustrated in Figure 5.1. With a discrete sample input, frequency inversion can be realised by simply inverting the sign of every other sample [Kak, *et al.*(1977)]. The frequency inversion is a one-to-one phoneme mapping process that can be learned [Blessner,(1972)]. Even in the absence of formal learning, the residual intelligibility of frequency inversion is non-zero. Speech sounds that are rich in tones near the centre of the speech band (1500 Hz in Figure 5.1) are subject to very small absolute values of frequency shift, and the result is a sound with a relatively small perceptual distance from the original sound [Jayant, *et al.*(1981)].

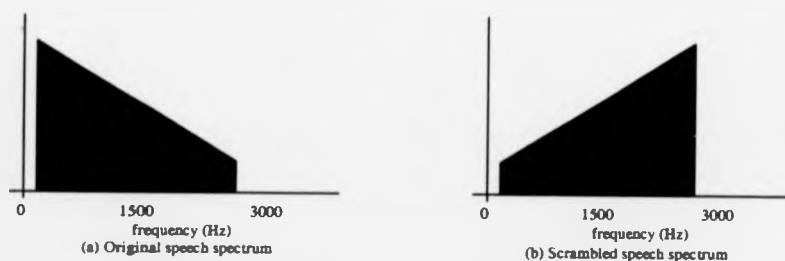


Figure 5.1: Frequency inversion

**Bandsplitting** [Nelson,(1976)]: In bandsplitting, the speech spectra is divided into  $n$  sub-bands. These sub-bands are then permuted in one or some of  $n!$  possible ways, which is given to both transmitter and intended receiver. The permutation order can be changed periodically according to an arrangement between the transmitter and the intended receiver. Like frequency

inversion, bandsplitting is an operation that is learnable, the correct position of 1 or 2 of  $n$  sub-bands often adequate to recognize input phonemes [McCalmont,(1974)]. In the example shown in Figure 5.2, the speech is split into 5 sub-bands, and the scrambler permutes the order of these sub-bands. In a dynamic bandsplitting system, the permutation algorithm is changed several times per second [Jayant,(1982)].

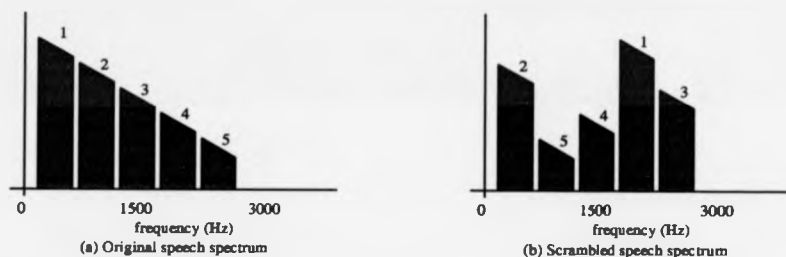


Figure 5.2: Bandsplitting

*Bandsplitting with Frequency Inversion* [Nelson,(1976)]: This technique is realised by combining the bandsplitting with local frequency inversion in time-varying selected sub-sets of sub-bands. This is illustrated in Figure 5.3. As in the case of frequency inversion, the residual intelligibility is high with trained listeners, producing word intelligibility scores in the range 45 to 70% [Kahn,(1967)]. In the example of Figure 5.3, the bandsplitting operation of Figure 5.2 is followed by local frequency inversion of 3 out of 5 sub-bands. In a dynamic system, the permutation algorithm, the number of inverted bands and the selection of the sub-bands for inversion, are all variables that are changed several times per second [Nelson,(1976)].

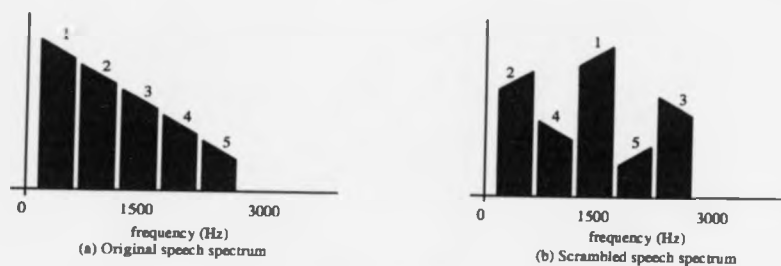


Figure 5.3: Bandsplitting combined with frequency inversion

*Frequency Inversion Followed by a Cyclic Bandshift* [Brunner,(1980)]:

This technique, shown in Figure 5.4, is similar to the procedure of bandsplitting with frequency inversion. The speech spectra is first inverted and then divided into  $n$  sub-bands. These sub-bands are frequency shifted in a cyclic manner several times per second. In the snapshot of Figure 5.4 this shift equals the width of two sub-bands. In a system with  $n = 16$  sub-bands, and shift variation of 50 per second, residual intelligibility scores were about 55% for separately spoken digits and about 30% for spoken words [Brunner,(1980)].

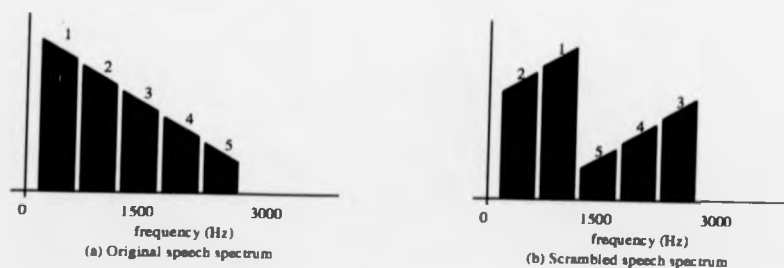


Figure 5.4: Frequency inversion followed by cyclic bandshift



### Time-Domain Scrambling

*Time Inversion* [Belland, *et al.*(1978)]: The time inversion consists of inverting the order of speech samples, locally within frames, with each frame has duration between 128 and 256 ms. This scrambling process can result in substantial losses of intelligibility [Belland, *et al.*(1978)]. Figure 5.5 shows a example of the time inversion. The frame size in this particular case is 2048 samples, with 8000 Hz sampling rate, this represents a length of 256 ms. The weakness of this technique is that it has very little cryptanalytical value due to the simple inversion process.

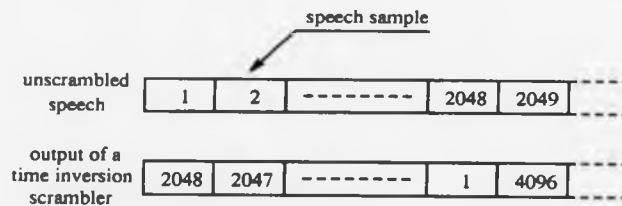


Figure 5.5: Time inversion

*Time Segment Permutation (TSP)* [Hong, *et al.*(1981)]: A time segment permutation is a time-scrambling procedure where segments of speech are permuted in pseudo-random fashion (Figure 5.6), with a segment-mapping algorithm known to both scrambler and descrambler. The choice of segment duration should reflect a compromise between the conflicting requirement of total communication delay (an increasing function of segment duration), and bandwidth expansion (a decreasing function of segment duration). A good

compromise is a segment that is 16 to 32 ms long [Jayant, *et al.*(1981)]. In the example of Figure 5.6, the TSP scrambler has a memory of  $b$  segments (here  $b = 8$ ). The scrambler outputs all the  $b$  segments of memory in a random order before refilling its memory with  $b$  incoming speech segments. The TSP scrambler has a total communication delay of  $2b$ . With  $b = 8$  and 16 ms segment duration, this delay would be 256 ms. The choice of  $b = 16$  provides lesser residual intelligibility, but with the greater delay of 512 ms [Jayant, *et al.*(1981)]. The TSP can be operated in two different manners, the block TSP [Leitich,(1977)], also called the hopping-window TSP, and sequential TSP [Jayant, *et al.*(1983)], sometime called the sliding-window TSP.

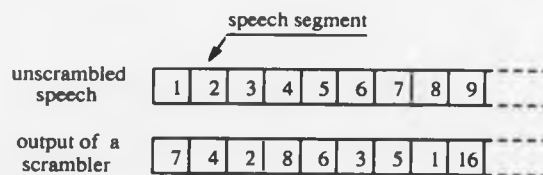


Figure 5.6: Time segment permutation (TSP)

The block TSP scrambler considers speech segments of typical duration 16 ms in blocks of  $b$  segments each, and completes the transmission of those  $b$  segments in random order before proceeding on to the next block of  $b$  segments.

The sequential TSP scrambler also begins with a memory of  $b$  segments of typical duration 16 ms, but it brings in a new incoming segment into mem-

ory as soon as one out of  $b$  contents of its memory is randomly selected and outputted for transmission. The scrambler operates with the additional constraint that if a segment stays in memory for a period of  $t$  segment durations, such a segment is then immediately outputted, regardless of the dictates of the pseudo-random segment selector. The special design of a staying time  $t = 2b$  has been noted to be optimal from the residual intelligibility point of view [Jayant, *et al.*(1983)]

### 5.2.2 Two Dimensional Speech Scramblers

One dimensional scramblers are characterised by significant amounts of residual intelligibility whether they use time domain manipulation or frequency domain manipulation [Jayant, *et al.*(1983)]. Two dimensional scramblers perform in both time and frequency domains to achieve an intelligibility which is often lower than that provided by one dimensional. Many algorithms were developed for the two dimensional scrambling, such as Band-splitting combined TSP [Kahn,(1967)] , Frequency Inversion combined with Block TSP [Jayant, *et al.*(1981)] and Frequency Inversion and Cyclic Band-shift combined with Time manipulation [Brunner,(1980)]. They are straight forward combinational operations of time and frequency scramblers. Time-Frequency Segment Permutation, or TFSP [Jayant, *et al.*(1983)], is illustrated in Figure 5.7. The scrambler memory consists of a matrix of  $bn$  time-frequency segments, generated by  $n$  sub-bands from each of  $b$  time segments. Time-frequency segments are outputted randomly in a way very similar to

the operation of sequential TSP, with the constraint that no segment stays in scrambler memory for more than  $t_f = 2bn$  segments duration. Every set of  $n$  contiguous time-segments is reconstituted into one pseudo-speech time segment for transmission. With the  $t_f = 2bn$  design, the total communication delay in the system is  $2b$ , the same as that of the TSP scrambler with a memory of  $b$  time segments. In fact, one dimensional TSP is the special case of  $n = 1$  in Figure 5.7.

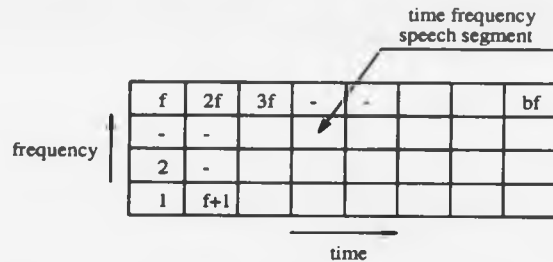


Figure 5.7: Time frequency segment permutation (TFSP)

### 5.3 Speech Scrambling Using Adaptive-Rate Sampling Algorithm

The optimal sampling rate algorithm for multi-band-pass signals developed in Chapter 3 has many applications in addition to that of speech compression. By using this algorithm, a sampling rate of  $\rho$  guarantees that all the translates by multiples of  $\rho$  of the spectral bands along the frequency axis are disjointed. If a signal is sampled at the rate  $\rho$  and then filtered in the usual way, the signal

obtained is that generated by those translates of the original bands or their reflections which lie in the range of the filter. The spectrum recombination caused by the sampling procedure can be used as frequency manipulation in signal processing. In this section, some examples are given to show the application of the sub-Nyquist sampling procedure in frequency manipulation with an emphasis on an application in speech scrambling.

*Spectral Shifting and Inversion:* This procedure can be employed to translate given elements of a specified spectrum to different ranges of frequency, without changing the spectral width of the elements. The procedure also allows given elements to be frequency inverted in addition to being frequency translated. It should be noted that the appropriate sampling rate will enable a single sampling process to perform multiple mixing functions simultaneously. Figure 5.8 illustrates an example of the procedure of spectra shifting and inverting using a sub-Nyquist sampling process. The original spectra contains two bands, with band *A* from 33500 to 36500 Hz and band *B* from 43500 to 46500 Hz (Figure 5.8(a)). It can be reconstructed with band *B* inverted by sampling at 16600 Hz and low-pass filtered at a cutoff frequency 6300 Hz (Figure 5.8(b)).

*Spectral re-ordering and Translation:* This procedure enables the elements of a given spectrum to be re-ordered and translated in frequency. Figure 5.9 is an example of such processing. A signal comprises a spectra of three sub-bands, band *A* from 32000 to 33000 Hz, band *B* from 36000 to 37000 Hz and band *C* from 43000 to 44000 Hz (Figure 5.9(a)). It can be reconstructed and

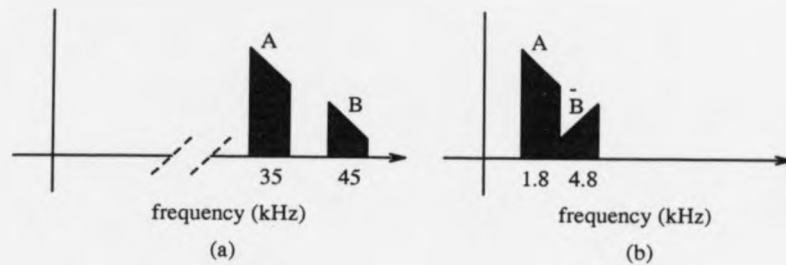


Figure 5.8: Spectral shifting and inversion

re-ordered by sampling at 6000 Hz and low-pass filtered at a cutoff frequency 3000 Hz (Figure 5.9(b)).

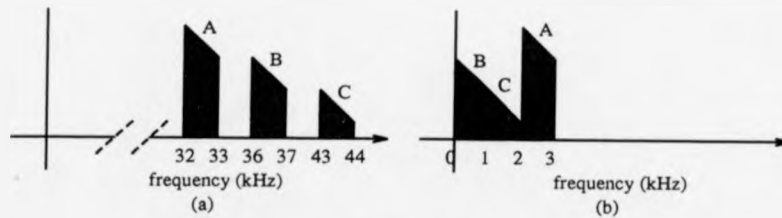


Figure 5.9: Spectral re-ordering and translation

Another example is given in Figure 5.10. Here the original spectrum contains three sub-bands of *A* from 32000 to 33000 Hz, *B* from 35000 to 36000 Hz and *C* from 41000 to 42000 Hz. Different sampling rates are used to re-order these bands. It can be seen from the reconstructed spectrum that the re-ordering in the fundamental interval can be achieved by using an appropriate sampling rate, but the bands can be widely separated and reflected bands can appear in the gaps.

*Spectral Scrambling and Recombination:* This procedure enables the sep-

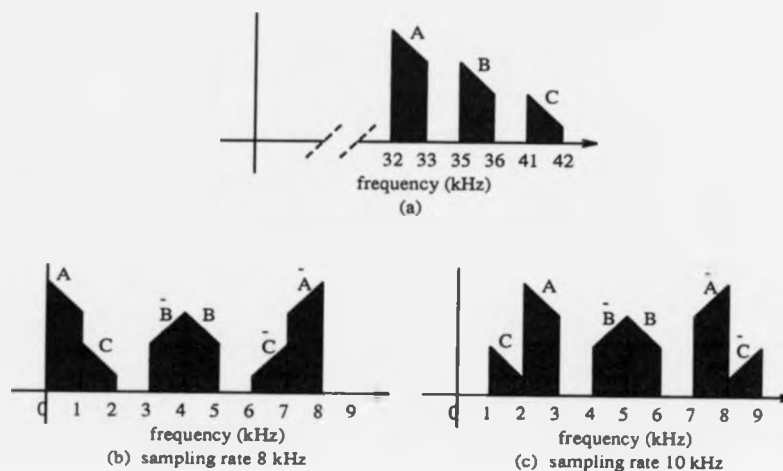


Figure 5.10: Spectral re-ordering and translation at different sampling rate

arated band-pass elements of a frequency scrambled signal, some of which may also be frequency inverted, to be recombined in their correct order, and possibly translated in frequency. This process enables the scrambled band-pass elements of a speech signal to be recombined into an intelligible speech signal. It thus provides the basis of a speech or data privacy/security communication system, in which the secure 'key' information comprises

- (a) the spectral frequency ranges of the scrambled elements
- (b) the unique value of sampling rate which will allow correct recombination.

If desired, dummy noise band-pass elements can be interleaved with the true wanted signal elements for additional security.

In the scrambling procedure, the baseband signal spectrum is divided into

several sub-bands. These sub-bands are then shifted to the specially arranged frequency ranges. The order of the original signal bands are changed and some bands may be frequency inverted. At the receiver, sampling the scrambled signal at a correct sampling frequency will recombine the sub-bands into their correct order. As an example, consider a signal with bandwidth of 4000 Hz, with 4 sub-bands defined by:  $A=[0,1000]$ ,  $B=[1000,2000]$ ,  $C=[2000,3000]$  and  $D=[3000,4000]$  Hz respectively, as shown in Figure 5.11(a). These sub-bands are then shifted according to the arrangement shown in Figure 5.11(b), i.e.

A at [8000,9000] (i.e. with carrier frequency 8500 Hz)

B at [17000,18000] (i.e. with carrier frequency 17500 Hz)

C at [18000,19000] (i.e. with carrier frequency 18500 Hz)

D at [11000,12000] (i.e. with carrier frequency 11500 Hz)

A unique sampling frequency can be calculated by using the method outlined in equations 3.6-3.9; in this case, it is 8000 Hz. Sampling the scrambled signal at 8000 Hz, and filtering with a 4000 Hz cutoff low-pass filter, will result in the desired signal spectrum illustrated in Figure 5.11(c). The bar over the letter designating a sub-band indicates a spectral inversion. Sampling rates near 8000 Hz give a partially scrambled version of the original signal: for example, at 8500 Hz the overlapping spectrum of Figure 5.11(d) results.



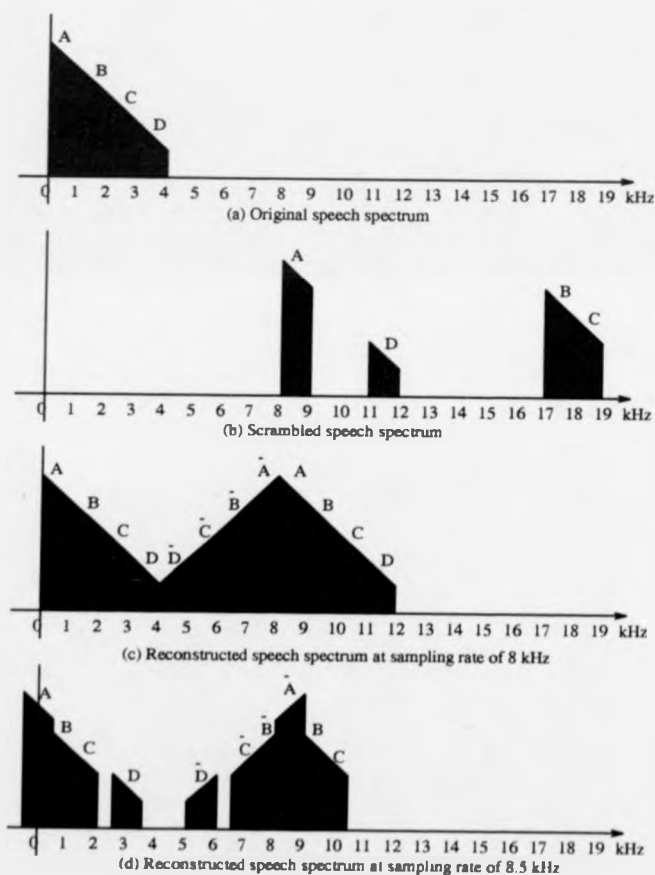


Figure 5.11: Example of the scrambling procedure

## 5.4 Scrambling System Description

The degree of security provided by a speech scrambling system is related to

- (1) the amount of intelligibility left over in the scrambled signal (residual intelligibility) and
- (2) the number of keys available for scrambling and descrambling.

The new speech scrambling algorithm proposed in this chapter is basically a frequency domain speech scrambling method. Similar to conventional frequency domain scrambling techniques described earlier in this chapter, it uses the knowledge of original speech spectra as one of the keys for descrambling. But in order to reconstruct the original speech, an additional key is needed, which is the reconstruction sampling frequency. Only the user who knows both keys is able to reconstruct the speech from the scrambled signal. This makes the new technique more secure than normal frequency domain scrambling techniques.

Figure 5.12 indicates the block diagram of the proposed scrambler and descrambler. The input speech is sampled and digitised. An FFT operation is then taken to transform the speech samples into frequency responses. The spectra is divided into  $n$  sub-bands which are spectrally shifted and recombined, possibly with frequency inversion on some of the sub-bands. To achieve a high security effect, a dummy noise spectrum can be inserted into the gaps between the scrambled sub-bands to mask the speech signal. The

scrambled signal together with the added noise is transformed into a time domain waveform and converted back to an analog signal for transmission. At the receiver, the transmitted signal is filtered by a multi-band-pass filter to remove the added noise components. It is then sampled at a unique sampling frequency which allows a re-ordering of the scrambled spectra to achieve the correct one. Finally, the reconstructed speech is converted back to an analog waveform and sent to its destination.

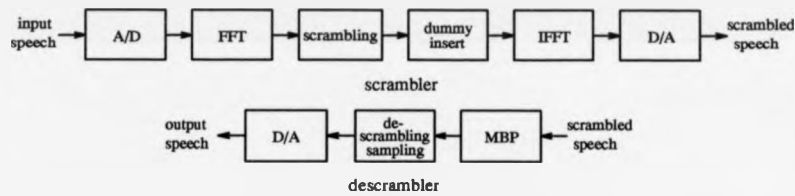


Figure 5.12: Block diagram of scrambling system

In this technique, the scrambled signal has a much wider bandwidth than that for the original speech. This bandwidth expansion makes it possible for several users to share the same communication channel, so this scrambling technique can be used for both single user and multi users.

#### 5.4.1 Single User System

In the single user case, a frame of speech is scrambled in one of the possible scrambling patterns. The speech bandwidth is expanded from its original 4000 Hz to 20000 Hz (in the example of Figure 5.11). This results in some significant gaps between the sub-bands. The existence of these gaps satisfies

the condition of employing the optimal sampling algorithm for multi-band-pass signals, which was developed in Chapter 3.

As mentioned earlier, the pattern of the scrambled spectrum and the descrambling sampling frequency are the 'key' factors to reconstruct the speech. Clearly, if only one scrambling pattern is used, there is only one descrambling sampling rate available. An unauthorised user can easily find the correct descrambling sampling rate by adjusting the sampling frequency on his receiver. Therefore, the reconstruction sampling frequency must be regularly changed, which requires change of the scrambling spectrum pattern. In practice, a number of available scrambling patterns can be stored. Figure 5.13 shows an example of these patterns. As can be seen, the different spectrum arrangements result in different descrambling sampling frequencies. A pseudo-random sequence generator can be used at the transmitter to decide on the scrambling pattern and descrambling sampling frequency for each frame of speech. At the receiver, the same sequence generator is employed to reconstruct the speech. Communication privacy can be established by distributing the code to only the authorised users.

#### 5.4.2 Multi-User System

The communications resource can be used more efficiently by allowing several users to share the system bandwidth. In Figure 5.14, speech from the source  $i$ , where  $i = 1, 2, \dots$ , is spectrally scrambled to produce a unique spectral arrangement. The resulting scrambled speech is then transmitted over the

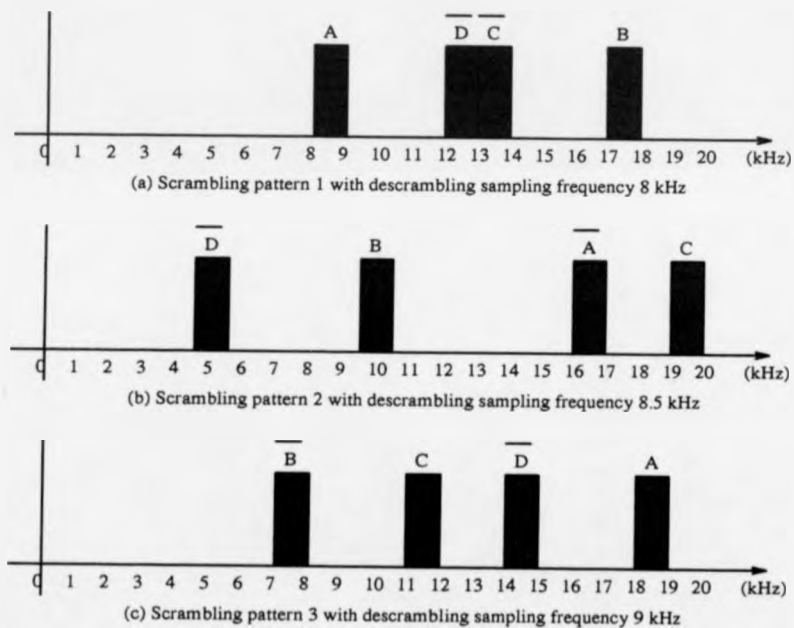


Figure 5.13: Scrambled spectrum arrangement (with system bandwidth of 20 kHz)

channel where it combines with the other scrambled speech signals to produce a composite scrambled signal. In the multi-user system, signals other than the desired signal will be treated as the unwanted signals. At the receiver, the unwanted bands will be filtered out and only the desired bands will be kept. A reconstruction procedure then takes place which involves sampling the received scrambled speech at the unique rate to reorder the scrambled bands back into the original format.

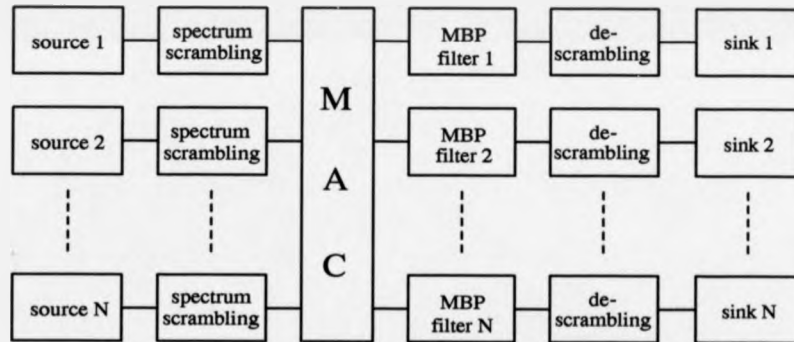


Figure 5.14: Multi user speech scrambling system

In such a system, constraints are imposed upon the scrambling process since it is essential to avoid any band overlap among the different signals. Ideally, to avoid interference between any of these bands, and to ease filtering problems, bands should be as far apart as practicable. In the single-user case, the scrambled signal band format may be changed from time to time to prevent any unauthorized reconstruction. In the multi-user system at any particular time, this flexibility is reduced because of the presence of the other

scrambled signals; thus manipulation of scrambled bands must be carried out in a co-ordinated manner. At any time, each signal has its own unique scrambled band format with a corresponding unique reconstruction sampling rate. An example of a 3-user system is now presented. The arrangements for these three scrambling patterns are shown in Figure 5.13. The reconstruction sampling rates for the 3 signals are 8000, 8500 and 9000 Hz respectively. Figure 5.15 shows how the scrambling pattern of the three signals to be transmitted may be changed with time within the overall transmission bandwidth of 20000 Hz; each of the 3 input signals is assigned a different spectral format in each frame according to a predefined sequence. Consequently, the sampling rate used by the individual receiver will also have to change from frame to frame.

The multiple access system is, in effect, a type of Code Division Multiple Access (CDMA) involving a hybrid combination of Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA) [Sklar,(1988)] However, in contrast to conventional CDMA, the frequency range allocated to each user does not cover the entire transmission bandwidth. Because of the sub-sampling procedure at the receiver, there is a reconstruction sampling rate tolerance range. Sampling the scrambled signal at any sampling rate falling in this range will yield the desired signal. In order to avoid sampling rate ambiguities among the users, the sampling rate tolerance ranges for all the users have to be disjoint. The implementation of the multi-user speech scrambling system is similar to the single user sys-

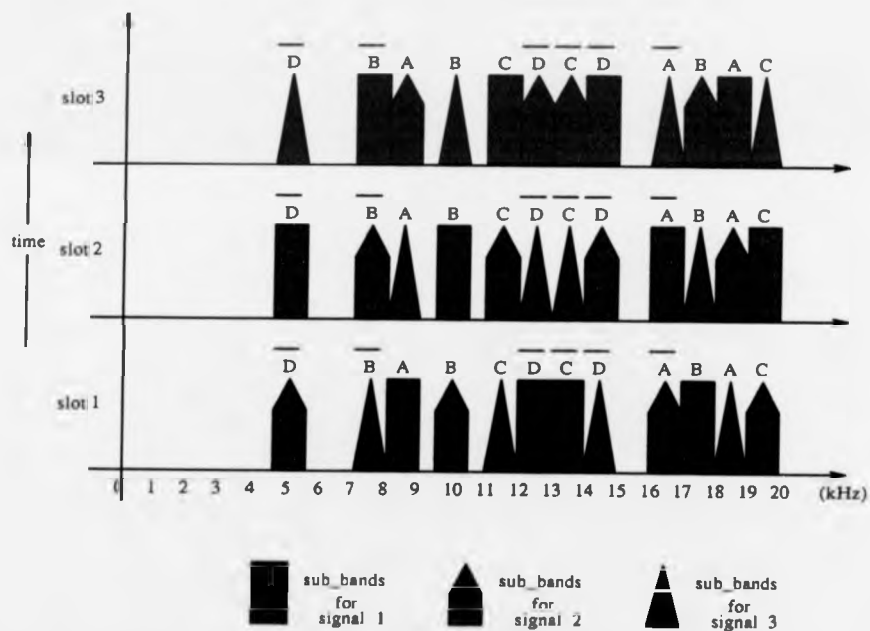


Figure 5.15: Communications resource plane in three users system



tem. A number of available transmitted signal formats can be stored. A pseudo-random sequence generator can be used at transmitter to select the scrambled spectrum for each user at each frame of speech. It is essential that this selection of scrambled spectrum formats will not cause any spectral overlap between each user. At each receiver, a synchronised version of the same sequence is employed to select the appropriate sampling frequency and hence reconstruct the speech waveform. On one hand, more sub-bands provide for better scrambling; on the other hand, however, more sub-bands complicate the receiver processing.

### 5.4.3 Sampling Rate Tolerance

Theoretically, only one sampling frequency can be used to reconstruct the signal exactly. The reconstruction sampling involves taking a partial set of samples from the received samples. Therefore, there is a tolerance range for the reconstruction sampling frequency which is dependent upon the system bandwidth,  $W_s$ , and the message signal bandwidth,  $W_m$ . The input signal is firstly sampled at  $\rho_s = 2 \times W_s$  Hz, then low-pass filtered with cutoff  $W_m$  Hz, where  $W_m < W_s$ . The scrambled signal has a spectra in the range of  $(0, W_s)$ . Suppose, the reconstruction sampling rate is  $\rho_m$ ; the reconstruction sub-sampling will therefore take one sample from every  $\lfloor \rho_s / \rho_m \rfloor$  received samples, where  $\lfloor x \rfloor$  denotes the nearest integer to  $x$ . There is more than one sampling frequency which can take the same samples from the received signal. If  $\rho'_m$  is one of the possible sampling rates different from  $\rho_m$ , but taking the

same samples from the scrambled signal, it must satisfy the condition

$$\left| \frac{\rho_s}{\rho'_m} - \frac{\rho_s}{\rho_m} \right| < 0.5 \quad (5.1)$$

Since  $\rho_s = 2 \times W_s$

$$\left| \frac{W_s}{\rho'_m} - \frac{W_s}{\rho_m} \right| < \frac{1}{4} \quad (5.2)$$

From 5.2

$$\frac{\rho_m W_s}{(W_s + \frac{1}{4}\rho_m)} < \rho'_m < \frac{\rho_m W_s}{(W_s - \frac{1}{4}\rho_m)} \quad (5.3)$$

The percentage tolerance of sampling frequency is defined by

$$\frac{|\rho'_m - \rho_m|}{\rho_m} \times 100\% \quad (5.4)$$

From 5.3 and 5.4, it can be seen that the percentage sampling frequency tolerance is dependent on both  $W_s$  and  $\rho_m$ .

Figure 5.16 shows the relationship between percentage sampling frequency tolerance, system bandwidth  $W_s$  and the theoretical reconstruction sampling frequency  $\rho_m$ .

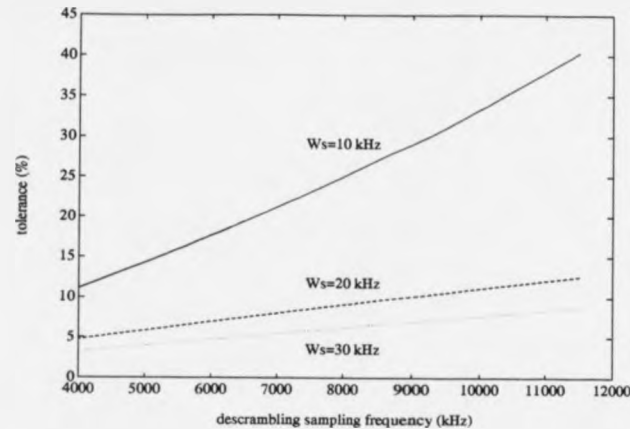


Figure 5.16: Descrambling sampling frequency tolerance for various system bandwidth

## 5.5 Subjective Tests to Measure Residual Intelligibility

The residual intelligibility was measured by using subject intelligibility test. The speech used for this test contains twenty-five sentences, one hundred individual words and one hundred 4-digit numbers, such as 4983, read as four-nine-eight-three. They were read by an adult male and recorded on audio tape in a virtually noise free room. The scrambled speech was recorded on a tape and played back to listeners. Four untrained listeners took part in the experiment. They were asked to listen to the scrambled speech and try as hard as possible to guess the context of the speech. The average correct guesses from all the listeners were defined as the residual intelligibility.

The speech spectra can be divided into different numbers of sub-bands

to achieve different scrambling effects. The number of sub-bands in the tests were 4, 8 and 16 respectively. In each case, three available scrambling patterns were chosen and they are illustrated in Figure 5.13, 5.17 and 5.18 respectively.

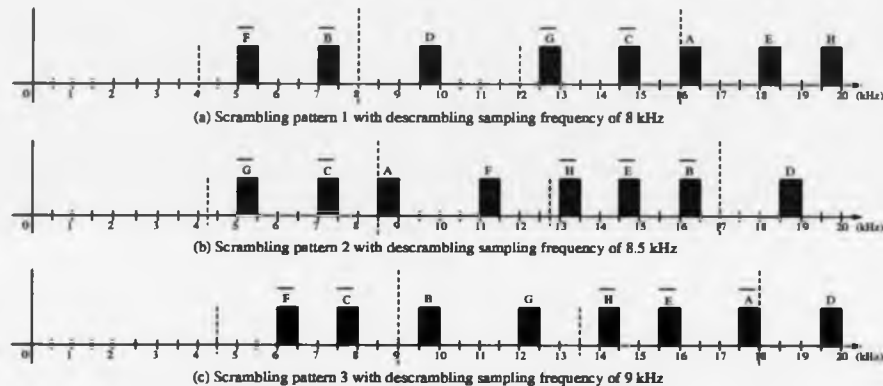


Figure 5.17: Three scrambling patterns for 8 sub-bands

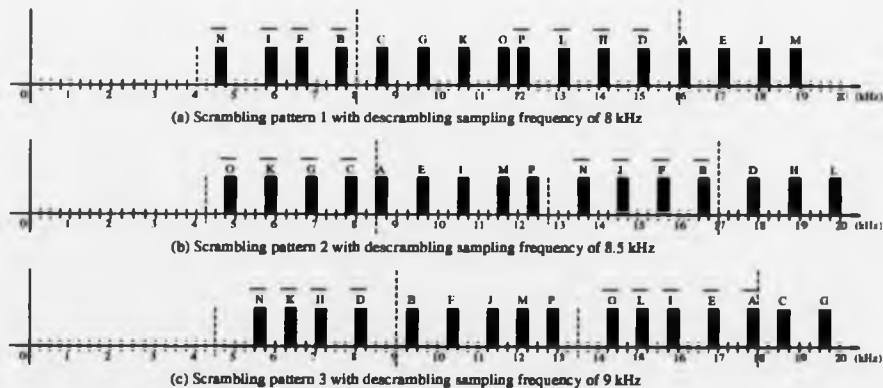


Figure 5.18: Three scrambling patterns for 16 sub-bands

The first set of tests was carried out for fixed pattern scrambling, i.e. dur-

ing each test, only one of the available scrambling patterns was employed. The results show that the scrambled speech by using both 8 sub-bands scrambling and 16 sub-bands scrambling have intelligibility of virtually zero. All four listeners failed to write down any of the spoken sentences, words and digits. In the case of 4 sub-bands scrambling, part of the speech rhythm was left in the scrambled speech and sometimes a correct guess could be made. This usually happens to the spoken digits due to the limited available choice (from 0 to 9). The test results show that the residual intelligibility for 4 sub-bands scrambling is about 5 per cent and all the correctly detected speech occurs for spoken digits.

The second set of tests was carried out for dynamic pattern scrambling. In the dynamic pattern scrambling, the speech spectrum were scrambled in different patterns at different times. In this test, the available scrambling patterns were stored. A pseudo-random generator was used to decide which pattern was used for each frame of speech. The pattern changed for each frame at a rate of 32 ms. The residual intelligibility for 4 sub-bands scrambling was reduced to zero by using the dynamic scrambling. This result shows that the changing of scrambling pattern does not only protect the speech from unauthorised descrambling but also helps in reducing the residual intelligibility.

## 5.6 Simulation Results

Computer simulation was carried out to verify the proposed scrambling techniques. In the simulation, only 4 sub-band scrambling was employed due to its low residual intelligibility and relatively simple procedure. Although the 8 sub-band and 16 sub-band scrambling also produce extremely low residual intelligibilities, they are more complicate than the 4 sub-band scrambling.

The speech signal was A/D converted by using a digital signal processor. The speech samples were then fed into the scrambler in frames, with each frame containing a 32 ms signal. An FFT was taken for each of these frames and the resulting frequency responses were divided into 4 sub-bands with each occupying a frequency range of 1000 Hz. The spectra was scrambled in the defined manner. In this simulation, three scrambling patterns were used (Figure 5.13). A pseudo random sequence was employed to choose one of the the patterns for each of the frames. As the frame size was chosen as 32 ms, the changing rate of the scrambling pattern, and consequently the descrambling sampling frequency, was every 32 ms.

Figure 5.19 shows the simulation result of the 4 sub-band scrambling technique. The figures include the original speech which is a 4-digit number, 4123, read by an adult male, scrambled speech, reconstructed speech at an incorrect sampling frequency and reconstructed speech at the correct sampling frequency. To illustrate the result more clearly, details of the sections of waveforms (for the word of 'two') are displayed in Figure 5.20.

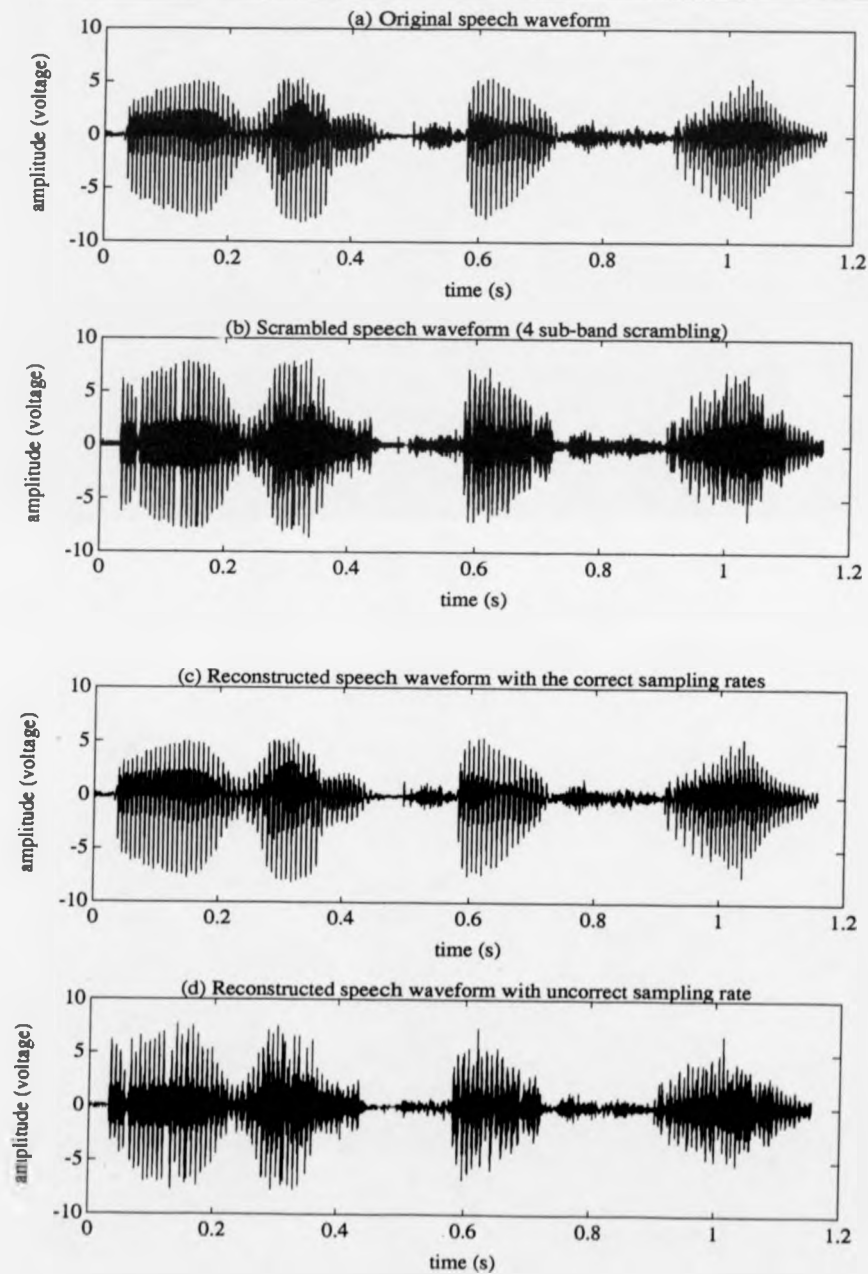


Figure 5.19: A simulation result of 4 sub-band scrambling technique

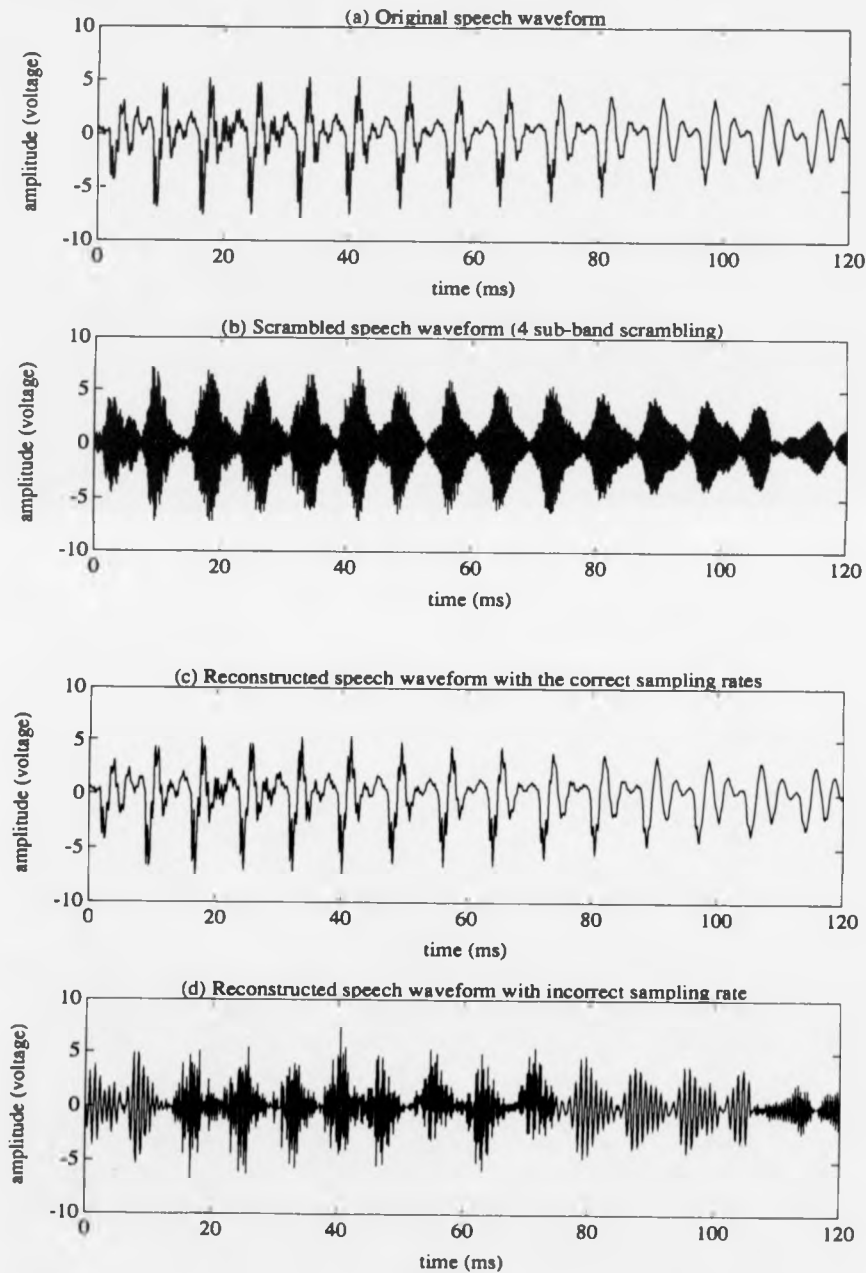


Figure 5.20: A detailed illustration of the result of the scrambling technique



## **Chapter 6**

# **Experimental Results and Practical Implementations**

## 6.1 Introduction

The adaptive-rate sampling (ARS) techniques developed in this thesis have great potential in low rate speech transmission application. When combined with low bit rate speech coding techniques, it produces better speech quality than that of low rate sampling techniques or low bit rate coding techniques at same transmission rate. Speech transmission reliability can be improved by combining the adaptive-rate sampling algorithm with error control coding, such as convolutional coding. This chapter presents some examples of the application and the experimental results. Also in this chapter, is an example of a practical implementation of the adaptive-rate speech technique, using Digital Signal Processor.

The speech samples used in the tests were read by two adult males and two adult females. There were two sets of speech, one contained 80 sentences and another contained 200 words (see the Appendix) which were carefully chosen by considering phonetic balance, word length, stress position, word importance, etc. The total duration are 4 minutes and 48 seconds for the sentences and 4 minutes and 32 seconds for words. The speech was first sampled at Nyquist rate of 8000 Hz and encoded with 8-bit PCM, using an A/D converter on a digital signal processor (DSP) board. The digitised speech was then processed in the speech compression systems and the reconstructed speech converted back to an analogue waveform and recorded on an audio tape, or played back over a loudspeaker.

## 6.2 Combine ARS with Speech Coding

The purpose of reducing the number of speech samples in digital speech transmission is to achieve a low transmission rate so that the speech can be transmitted safely over a channel which has a limited transmission capacity. As the transmission rate reduction can be realised both by low rate sampling and by low bit rate speech coding, a combination of these two methods can reduce the transmission rate significantly further. In this section, two examples show such an application. They are (i) a combination of low rate sampling with Hadamard coding and (ii) a combination of low rate sampling with Adaptive PCM.

A question might arise here. What is the reason for reducing the transmission rate using combined low rate sampling and low bit rate speech coding, rather than reducing either the sampling rate or the number of encoded bits alone? To answer this question, let us examine two different cases. In the first case, the sampling rate or number of encode bits sometimes has already reached its minimum. For example, in 2-bit adaptive PCM [Wilkinson,(1973)], each speech sample is represented by 2 bits, one for polarity and another for amplitude. It is not possible to reduce the bit rate further. As the transmission rate is the number of encoded bits multiplied by the sampling rate, the only way of reducing the transmission rate further, is to reduce the sampling rate. In another case, even the sampling rate or number of encoded bits is not at its lowest, it is not practical to reduce them

arbitrarily in order to bring the transmission rate down. Simply reducing the sampling rate or the number of encoded bits may cause significant degradation of speech quality. An experimental result is shown in this section to explain this problem.

### **6.2.1 Comparison of Combined Technique with Low Bit Rate PCM and Low Rate Sampling**

To illustrate the advantages of the combined technique, a number of tests were carried out to measure the reconstructed speech quality in three different cases, low bit rate PCM, low rate sampling and the combined technique. In the low bit rate PCM, the speech was sampled at the Nyquist rate of 8000 Hz and each sample was encoded with 2, 3, 4, 5, 6 or 7 bits respectively. At each level, the reconstructed speech quality was recorded and measured. In the low rate sampling technique, the number of encoded bits was fixed at 8. The sampling rate was reduced to 2000, 3000, 4000, 5000, 6000 or 7000 Hz to achieve low transmission rate. The results were used to make the comparison. In the combined technique, both the sampling rate and the number of encoding bits could be changed and the total transmission rate was the sampling rate multiplied by the number of encoded bits.

Both subjective quality and objective quality were measured. The subjective measurement was carried out in the same manner as that described in Chapter 3. The objective measurements were carried out in both time and

frequency domain. The time domain measurement was defined as distortion between the input and the output speech waveform [Kitawaki, *et al.*(1982)]. Segmental Signal-to-Noise Ratio ( $SNR_{seg}$ ), which is signal to noise ratio in a short intervals of speech (the short interval is set to 16 ms in the tests), was calculated as follow:

$$SNR_{seg} = \frac{1}{M} \sum_{n=1}^M 10 \log_{10} \left( \frac{\sum_{n=1}^m \sqrt{[x(n)]^2}}{\sum_{n=1}^m \sqrt{[x(n) - \bar{x}(n)]^2}} \right) \quad (6.1)$$

where  $x(n)$  is the input speech samples,  $\bar{x}(n)$  is the output speech samples,  $m$  is the number of samples in one segment and  $M$  is the total number of segments.

The objective measurement in the frequency domain was defined as distortion between the input speech and the output speech spectrum [Kitawaki, *et al.*(1982)]. The Spectral Distortion (SD) measure is defined as

$$SD = \sqrt{\frac{1}{m} \sum_{n=1}^m \{S_x(n) - S_y(n)\}^2} \quad (6.2)$$

where  $S_x(n)$  is input speech logarithmic spectrum,  $S_y(n)$  is the output speech logarithmic spectrum,  $m$  is the number of samples.

The SD measurement shows the logarithmic spectral distortion between the input and the output speech spectrum computed by Fourier transform. The Fourier transform was computed by 126 point FFT (corresponding to 16 ms frame size). Both Spectral Distortion and Segmental Signal-to-Noise Ratio are reasonably well correlated with the subjective quality measurement.

[Kitawaki.(1988)] [Goodman, *et al.*(1983)] Figure 6.1 shows the subjective measurement (intelligibility test) for the low sampling rate system and low bit rate PCM system. It can be seen that at transmission 48 kb/s or less, the low rate sampling process produced better speech quality than that of low bit rate PCM did. For example, at a transmission rate of 16 kb/s, the low rate sampling produces a speech with an intelligibility of around 86%; in contrast, the low bit rate PCM only produces 13% intelligibility. Figure 6.2 and 6.3 show the spectral distortion and segmental signal-to-noise ratio measurements respectively. These two graphs also show that the low rate sampling has less spectral distortion and a higher segmental signal-to-noise ratio.

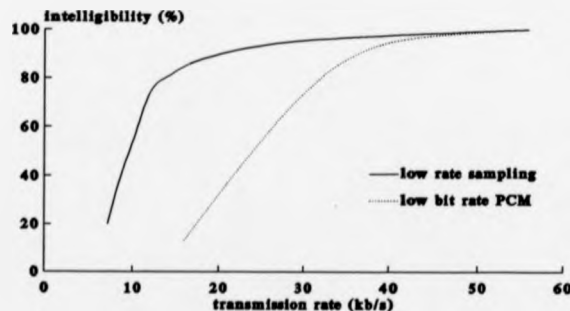


Figure 6.1: Intelligibility test for low rate sampling and low bit rate PCM

Simply reducing the number of encoding bits in PCM is not a practical method of bringing down the transmission rate. That is the reason why

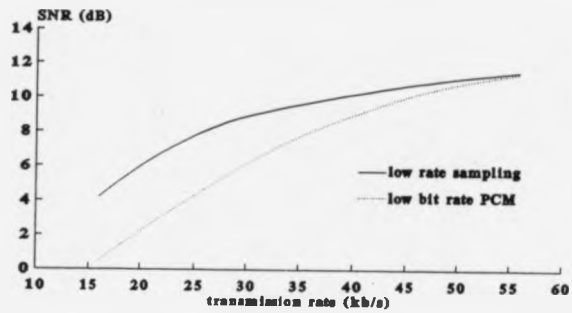


Figure 6.2: Segmental signal-to-noise ratio for low rate sampling and low bit rate PCM

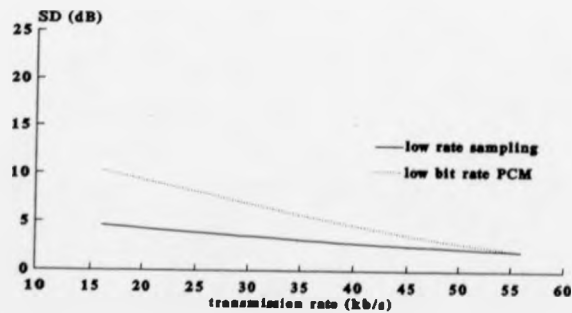


Figure 6.3: Spectral distortion for low rate sampling and low bit rate PCM

more sophisticated speech coding techniques are needed. The comparison made in this section shows that low rate sampling technique can produce highly intelligible speech at a relatively low transmission rate, even without using any low rate speech coding method.

A set of tests was conducted for the combined technique. The number of encoded bits was 7, 6, 5, 4, 3 or 2. At each encoding level, a different sampling rate was used to take the speech samples. Figure 6.4 shows the intelligibility test results. Compared with Figure 6.1, it can be seen that at the same transmission rate, the reconstructed speech quality in the combined system may be higher than that in either the low rate sampling or low bit rate PCM system. For example, 6-bit PCM at a sampling rate of 3000 Hz produces a speech with intelligibility of 90%. The corresponding transmission rate is 18 kb/s. For the low rate sampling alone to achieve the the same speech quality, the transmission rate should be around 22 kb/s (although the reduction is not significant, bearing in mind that a proper low rate speech coding was not used). The following section will illustrate the great potential of transmission rate reduction using combined low rate sampling and speech coding techniques. Figure 6.5 and 6.6 also show, from the spectral distortion and segmental signal to noise ratio points of view, that the combined system produced the best reconstructed speech quality.



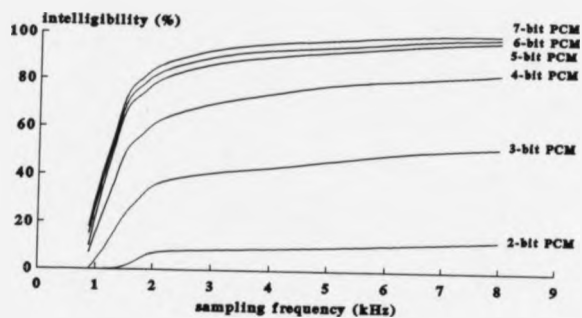


Figure 6.4: Intelligibility test for combined system

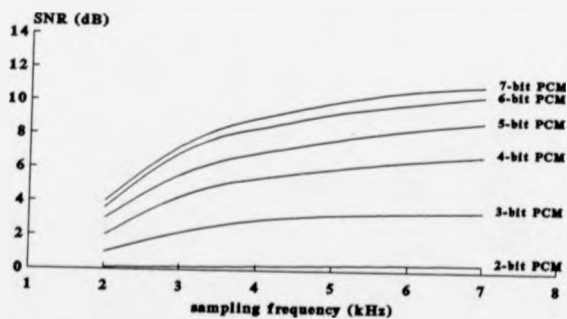


Figure 6.5: Segmental signal-to-noise ratio for combined system

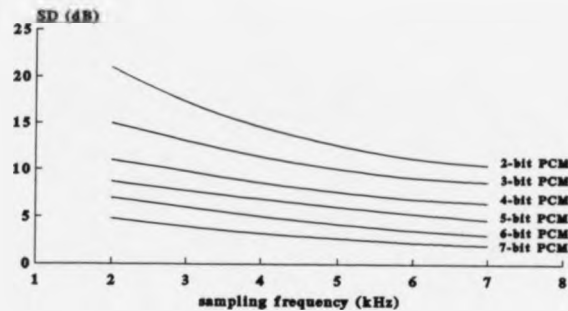


Figure 6.6: Spectral distortion for combined system

### 6.2.2 Combine Low Rate Sampling with Hadamard Coding

An orthogonal transformation - the Hadamard transformation [Shum.*et al.*(1973)] [Frangoulis,(1977)] - has been used to reduce the speech data rate after the low rate sampling process. Here, the samples of speech were analysed in terms of their Hadamard transform coefficients. Only the dominant coefficients were then transmitted over the communications link in quantised form; at the receiver, the received coefficients were used to reconstruct an estimate of original speech signal via the inverse Hadamard transformation.

### Hadamard Matrix and Hadamard Transform

The Hadamard matrix is a square array of plus and minus ones, whose rows and columns are orthogonal to each other. Hence, the product of the matrix and its transpose is the identity matrix multiplied by a constant  $N$ , where  $N$  is the order of matrix, and the lowest value it may assume is two, thus giving the lowest order Hadamard matrix as

$$H_2 = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} \quad (6.3)$$

By limiting  $N$  to an integral power of two, symmetrical Hadamard matrices may be obtained recursively by

$$H_{2N} = \begin{vmatrix} H_N & H_N \\ H_N & -H_N \end{vmatrix} \quad (6.4)$$

Each row of a Hadamard matrix is called a Hadamard function  $Had(j, k)$  for  $j = 0, 1, \dots, N - 1$ .

The one-dimensional discrete Hadamard transform of a real signal is defined as [Shum, *et al.* (1973)]

$$F(j) = \frac{1}{N} \sum_{k=0}^{N-1} f(k) \times Had(k, j) \quad j = 0, 1, \dots, N - 1. \quad (6.5)$$

where

$F(j)$  is the  $j$ th normalized Hadamard coefficient,  $f(k)$  is the  $k$ th discrete

sample of signal,  $Had(j, k)$  is the  $j$ th Hadamard function.

In this case, the Hadamard transform is called an  $N$ -point Hadamard transform.

The inverse transform is given by

$$f(k) = \sum_{j=0}^{N-1} F(j) \times Had(k, j) \quad k = 0, 1, \dots, N-1. \quad (6.6)$$

### Hadamard Transform Coding

In applying the Hadamard transformation technique to speech signals, the speech to be processed is divided into short-time segments with each contains  $N$  samples  $f(0), f(1), \dots, f(N-1)$ . The calculation of equation 6.5 is then performed to each of these frames to generate the  $N$  Hadamard coefficients  $F(0), F(1), \dots, F(N-1)$ . For example, if the frame size is set as 8 ms, each frame contains 64 samples at the Nyquist sampling rate of 8000 Hz. By using equation 6.5, 64 Hadamard coefficients can be obtained from 64 samples of speech.

Experimental results indicated that among the 64 Hadamard coefficients, between 60% and 90% of the total average power was concentrated in the 6 to 10 dominant coefficients [Frangoulis, (1977)]. Because of this fact, for speech transmission, it is possible to encode and transmit only the dominant coefficients instead of the original speech samples. Therefore, the potential for data compression exists. In the case of a 64-point Hadamard transformation, almost 90% of total power is concentrated in 10 dominant coefficients.

By transmitting only these 10 dominant coefficients, a reasonable quality of reconstructed speech can be achieved at the receiver. However, if the number of dominant coefficients transmitted is less than 5, the reconstructed speech will be of very low intelligibility and noisy. [Frangoulis,(1977)]

Because only a subset of the total coefficients is required to be transmitted, extra bits have to be used to indicate to the receiver the order of any particular received coefficient. In the case of the 64-point Hadamard transform, 6 bits are enough to identify the dominant coefficients. The transmission rate can then be calculated as following formula

$$w = L \times \frac{(m + l)}{N} \times R \quad (6.7)$$

where  $L$  is the number of coefficients used in reconstructing the signal (i.e. the dominant coefficients),  $R$  is the sampling rate,  $N$  is the number of samples in one frame,  $m$  is the number of bits used to encode each coefficient and  $l$  is the number of bits needed to indicate to the receiver which of  $N$  coefficients a particular received coefficient is.

Suppose 8-bit PCM is used to encode the dominant coefficients, the transmission rate is

$$w = 10 \times \frac{(8 + 6)}{64} \times 8000 = 17.5 \text{ kb/s} \quad (6.8)$$

However, the transmission rate is 64 kb/s when Hadamard coding is not used.

### Experimental Results

In order to use both the Hadamard transform and the fast Fourier transform, the sampling rates in the adaptive-rate sampling system were limited, and at each rate the the number of samples taken in one frame was an integral power of 2. Here six sampling rates are used: they are 250, 500, 1000, 2000, 4000 and 8000 Hz. The input speech signal was first sampled at the Nyquist rate (8000 Hz). The frame size was chosen as 16 ms. 128 samples, the number of samples in one frame after the sub-sampling were 4, 8, 16, 32, 64 and 128 corresponding to sampling rate of 250, 500, 1000, 2000, 4000 and 8000 respectively. To achieve a high quality speech reconstruction, 8-PCM was used to encode each Hadamard coefficient ( 7-bit PCM encoding can also be used, but the reconstructed speech is slightly worse [Frangoulis,(1977)]). Table 6.1 shows, at different sampling rates, the number of samples per frame ( $N$ ), the number of dominant coefficients ( $L$ ), the number of bits used to identify the dominant coefficients ( $l$ ) and the transmission rates ( $w$ ).

<i>sampling rate</i>	250	500	1000	2000	4000	8000
<i>no. of samples per frame (N)</i>	4	8	16	32	64	128
<i>no. of identifying bits (l)</i>	2	3	4	5	6	7
<i>no. of dominant coefficient (L)</i>	1	2	3	4	10	18
<i>transmission rate (w)</i>	625	1375	2250	3250	8750	16875

Table 6.1: The relationship of sampling rate  $w$ ,  $N$ ,  $L$  and  $l$  ( $m = 8$ )

As mentioned in Chapter 3, at an average sampling rate of 3500 Hz, the reconstructed speech has high intelligibility and low noise. For this reason,

the sampling rate in this test was fixed at 4000 Hz for high quality reconstruction. The frame size was set as 16 ms, 128 samples at the Nyquist rate of 8000 Hz. After the sub-sampling, each frame contained 64 samples. A 64-point Hadamard transformation was then taken to obtain the 64 coefficients. 10 coefficients with highest absolute value were chosen to be encoded and transmitted whilst the rest were ignored. The transmission rate according to Table 6.1, is 8750 b/s. The intelligibility in this particular case is around 91%. Without Hadamard coding, the transmission rate is 32000 b/s and the intelligibility is 96%. The transmission rate was reduced by 72% at a price of slight speech degradation. The intelligibility for the same Hadamard coding without the low rate sampling process is about 95%. The transmission rate is 17500 b/s.

### 6.2.3 Combine Low-Rate Sampling with Adaptive PCM

The low-rate sampling process can easily be combined with speech waveform coding to achieve a low transmission rate. An example is given in this section to illustrate the combined low-rate sampling with adaptive PCM (APCM).

The APCM [Jayant,(1974)] technique used here works with a basic uniform quantiser, but with variable step size. For each sample, the step size is modified, based on the knowledge of which quantiser slots were occupied by the previous sample. The quantiser has  $N$  quantisation levels ( $\pm N = 1, 2, \dots, 2^{B-1}$ ,  $B$  is the number of encoded bits). If the original speech samples are encoded with 8-bit PCM, then the initial step size ( $\Delta_0$ ) for the

$B$ -bit APCM is set as  $2^{8-B}$ . For the first input sample, the quantiser detects which of the quantisation slots (the interval between two quantisation levels) it falls into. The code word  $H_0$  for this slot is then transmitted. The number of code words is  $B - 1$ , in addition there is one bit to indicate the polarity of the sample. The output of the reconstruction is

$$y_0 = \Delta_0 \times H_0 \quad (6.9)$$

For the second input sample, the step size is modified by the knowledge of the code word  $H_0$ , and  $\Delta_0$

$$\Delta_1 = \Delta_0 \times M(H_0) \quad (6.10)$$

where  $M$  is a time-invariant function of the code word  $H$ .

The second sample is then reconstructed as

$$y_1 = \Delta_1 \times H_1 \quad (6.11)$$

The same procedure is carried on until all the speech samples are processed.

The key factor in this technique is the multiplier function  $M(H)$ . When it is properly designed, the adaptation procedure (equation 6.10) adjusts the step size, at every sample, to an updated estimate of speech signal variance. Table 6.2 lists the optimal step-size multipliers for PCM [Jayant, (1974)].



<i>CODER</i>	<i>PCM</i>		
B	2	3	4
$M_1$	0.60	0.85	0.80
$M_2$	2.20	1.00	0.80
$M_3$		1.00	0.85
$M_4$		1.50	0.80
$M_5$			1.20
$M_6$			1.60
$M_7$			2.00
$M_8$			2.40

Table 6.2: Step size multipliers for  $B = 2, 3$  and 4.

The number of bits used in the test were 2, 3 and 4 respectively. At each encoded rate, the speech was reconstructed at different sampling rates. At a typical sampling rate of 3500 Hz, the transmission rates are 7, 10.5 and 14 kb/s respectively. Figure 6.7 shows the intelligibility test results for the combined technique. Figures 6.8 and 6.9 show the segmental signal to noise ratio and spectrum distortion respectively.

The improvement of such a combined technique can be observed by comparing the measurements in this test with similar test results in other techniques. For example, in the combined low rate sampling and 4-bit APCM system, at a sampling rate of 2000 Hz, equivalent to transmission rate of 8 kb/s, the intelligibility is as high as 87%. Without using low rate sampling, a 3-bit APCM is needed to achieve an intelligibility of 89% and the transmission rate is 24 kb/s (see Figure 6.7). An enormous transmission rate reduction has been achieved with an insignificant degradation of speech intelligibility. Low rate sampling with normal 8-bit PCM would require a

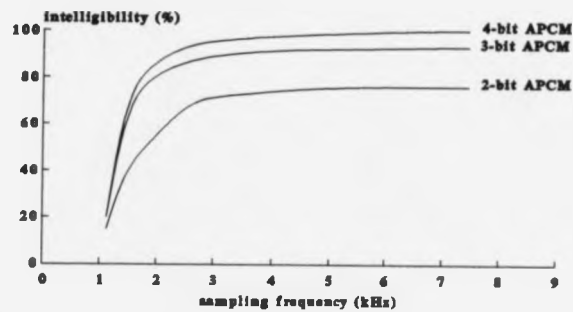


Figure 6.7: Intelligibility test for the combined low-rate sampling and APCM system

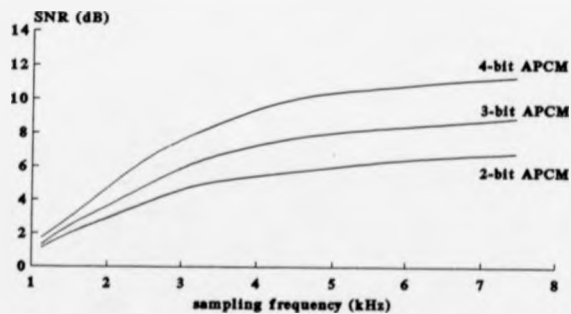


Figure 6.8: Segmental signal-to-noise ratio for the combined low-rate sampling and APCM system

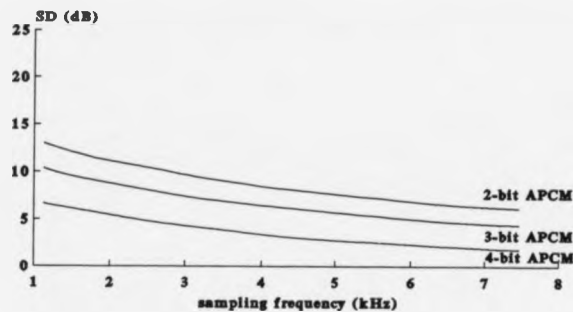


Figure 6.9: Spectrum distortion for the combined low-rate sampling and APCM system

transmission rate of 17 bk/s to achieve an intelligibility of 87% (see Figure 6.1).

### 6.3 Combine ARS and Channel Coding for Reliable Transmission

Any real communication channel has a limited transmission capacity. If the speech is transmitted at a rate which is much lower than the channel capacity, or say optimum transmission rate, the speech quality is impaired by unnecessary distortion of quantising or low rate sampling. If the speech is transmitted at a rate which is higher than the optimum rate, the transmission error can severely corrupt the transmitted information so that a

receiver fails to reconstruct the speech. In this section, an adaptive-bit-rate transmission is investigated, with the information rate adjusted according to the properties of the channel. The adaptation of sampling rate and transmission rate will always provide optimum speech quality for a prevailing channel condition. This requires the continuous monitoring of the channel state by Real-Time Channel Evaluation (RTCE) [Darnell.(1983)] [Honary, *et al.*(1988)] [Zolghadr, *et al.*(1988a)], which will adjust the adaptive rate speech system so as to be able to best cope with the limitations imposed by the current channel condition. To transmit the speech safely, a multi-level error control coding can be used, together with the adaptive sampling. This is to achieve a maximum reliability of speech transmission and optimal reconstruction of speech. A suitable coding scheme is chosen based on the knowledge of the condition of the transmission channel obtained from the RTCE monitor, which will guarantee a safety of transmission. The threshold level in the speech compression process is adjusted to reach a resulting sampling rate which takes the maximum number of samples allowed, so that an optimum reconstruction quality can be obtained for the available channel. For convenience, a constant signaling rate of 32-kb/s was used and the source rate and channel-coding rate were adjusted in response to channel conditions. During good channel conditions, all 32 kb/s were used for speech transmission. When the channel capacity was low, the source rate was reduced to either 24 kb/s or 16 kb/s and channel coding was introduced to control the distortion caused by the transmission error. The change of source

rate was realised by changing the sampling rate. Here 4-bit APCM was used to encode the speech samples and sampling frequencies were 8000, 6000 or 4000 Hz corresponding to the source rate of 32, 24 or 16 kb/s respectively.

The channel codes were convolutional codes [Elias,(1955)], with rates of  $1/2$  or  $3/4$ . At a rate of  $3/4$ , the 'punctured' code [Cain,*et al.*(1979)] realisation simplifies the decoder, because the encoder and decoder structures are the same as for the rate  $1/2$  code. In fact, the rate  $3/4$  punctured codes are rate  $1/2$  codes with a fraction (2 out of 6) of the channel bits deleted. Table 6.3 presents the transmission formats listed in order of increasing resistance to channel impairment.

	<i>Format 1</i>	<i>Format 2</i>	<i>Format 3</i>
<i>Sampling rate</i>	8000	6000	4000
<i>Source code</i>			
<i>bits/sample</i>	4	4	4
<i>kbits/second</i>	32	24	16
<i>bit/sample protected</i>	0	4	4
<i>Channel code rate</i>	<i>no code</i>	$3/4$	$1/2$

Table 6.3: Sampling rate, source and channel code formats.

### 6.3.1 System Description

The simplified block diagram of combined adaptive-rate sampling and channel coding is shown in Figure 6.10. The speech compression and expansion were discussed in detail in Chapter 3. The RTCE techniques will not be investigated in this work and are assumed to function satisfactorily.

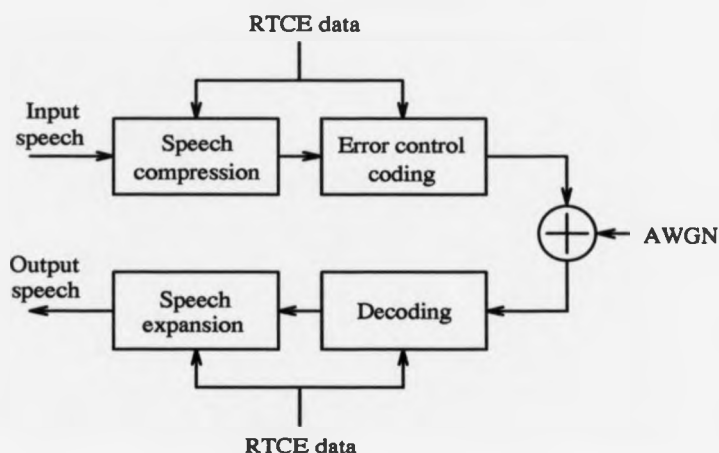


Figure 6.10: Simplified block diagram of combined adaptive-rate sampling and error control coding technique

The error control coding is a half rate convolutional encoder with a three-stage shift register, Figure 6.11. Each input bit is shifted into the leftmost stage and the bits in the register are shifted one position to the right. The output of the encoder consists of each modulo-2 adder (first from the upper adder, then from the lower one) to form the code symbol. For each input bit, there are two output bits, so that the ratio is  $1/2$ . The decoder uses the trellis algorithm [Viterbi, (1971)], which uses the trellis structure of the code and determines the maximum likelihood estimate of the transmitted sequences that has the smallest metric. The log likelihood function represents the metric which can be computed for each path in the trellis and is additive over the received sequence. The survivor is defined as the most probable path that has the smallest accumulated metric. With these definitions, the

Viterbi algorithm simply finds the path through the trellis with the smallest accumulated metric in such a way that it processes a received sequence in an iterative manner. At each step, it compares the metric of all paths entering each state, stores only the survivor with the smallest accumulated metric, and discards the unlikely paths at every state, which reduces the decoding effort. Therefore the Viterbi decoder must produce an estimate of the code sequence based on the received sequence.

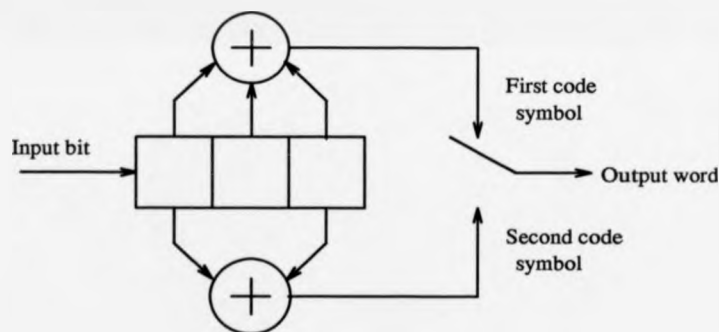


Figure 6.11: 1/2 rate convolutional coder ( $k=3$ )

At a rate of  $3/4$ , the decoding is significantly simplified by using punctured coding which periodically deletes bits from the  $1/2$  rate convolutional code. The optimal deleting map for the convolutional code in Figure 6.11 is shown in Figure 6.12 [Hole.(1988)]. Here, the output of the  $1/2$  rate convolutional encoder is divided into blocks of six bits. In each block, two out of six bits are deleted according to the deleting map. The remaining four bits correspond to three information bits, hence a convolutional code of rate  $3/4$  is generated. At a receiver, an inserter is used to restore the incoming

stream to their original length thus allowing the 1/2 rate decoder to be used. Since the nature of the deleted digits is not known at the receiver, erasures (unknown digits) are inserted in place of the deleted bits, using same deleting map used at the transmitter.

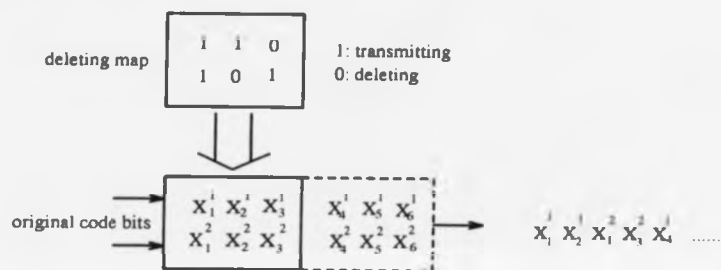


Figure 6.12: The deleting map for the rate of 3/4 convolutional code

### 6.3.2 Simulation Result

A computer simulation was carried out to verify the proposed system. The test was conducted at SNR ranges from 6 to 50 dB. At each level of SNR, all three formats of transmission were tested and the results were measured in the terms of segmental SNR. Figure 6.13 presents the results in the range of 6 to 16 dB of the channel SNR.

This figure shows when the channel is good enough (the channel signal-to-noise is higher than 13 dB) transmission format 1 provides the best performance with segmental signal-to-noise higher than 11 dB, and intelligibility of 100% (see Figure 6.7). In this range of channel SNR, the formats 2 and 3 reconstruct speech with segmental SNR of 9.5 and 11 dB, and intelligibility



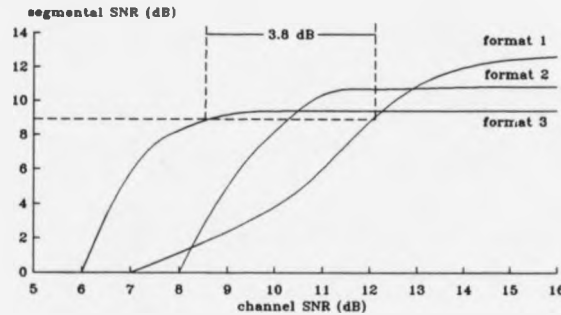


Figure 6.13: The performance of the three transmission formats as a function of channel SNR

of 97.5% and 99% respectively. The distortion was caused by the elimination of the trivial frequency components in the speech compression. From 10.5 to 13 dB of the channel SNR, the signal distortion in format 1 is excessive due to lack of error protection. In this range, transmission format 2 presents the highest transmission capability. The 3/4 rate convolutional coding provides adequate error protection to the transmitted data. The sampling rate in format 2 is 6000 Hz, this means three quarters of the total frequency components are preserved. Compared with the format 3 which deletes half of the total frequency components, the format 2 has less distortion from the speech compression. When the channel SNR is reduced to less than 10.5 dB, the channel distortion become so severe that the 3/4 rate convolutional coding in transmission format 2 fails to provide a useful error correction to overcome

the distortion. However the 1/2 rate convolutional coding in transmission format 3 is powerful enough to extend the channel viability during these conditions. Although, of the three transmission formats, format 3 exhibits the most serious distortion from the speech compression, it still provides reconstructed speech with an intelligibility of 97.5% when there is no channel distortion. As far as intelligibility is concerned, this is adequate for speech transmission. Assuming, a segmental SNR within 1 dB of the SNR of error-free format 3 transmission provides adequate speech quality, from Figure 6.13 it is estimated that relative to a conventional 32 kb/s transmission, adaptive-bit-rate operation extends the useful range during poor channel conditions by 3.8 dB.

## 6.4 Application of ARS in Speech Storage

When speech is needed to be stored for future use, the adaptive-rate sampling technique can be employed to make an efficient use of the storage space. Such an application is not difficult to implement. The analogue speech signal is first sampled at the Nyquist rate of 8000 Hz. The sampled speech is then processed in short-time frames, each of them lasting for 16 ms (containing 128 samples). Each frame of speech is analysed in frequency domain and redundancies are eliminated according to the requirements. Both Frequency Companding and Time Domain Compression techniques (see Chapter 3) can be used to compress the speech. The sub-sampling process takes a fraction

of samples from the original samples in each frame. Finally, the information bits are stored, along with the control bits for speech reconstruction. During the reconstruction process, the stored data are read from the store into the system. From the control bits, all the information needed for reconstruction, such as number of bits for each frame, spectrum structure etc, is known exactly. The reconstruction is then carried out and the resulting data is converted back into analogue speech and sent to its destination which could be a loudspeaker or telephone.

To illustrate the process described above, an example is given here for a segment of speech signal with a duration of 32 ms, or 256 samples. This segment is divided into two frames, each containing 128 samples. Figure 6.14 shows the speech waveform; the crosses on the waveform indicate the sampling points. The integral values of the samples are shown in Table 6.4.

The 128 samples in each of the frames were Fourier transformed to frequency domain. The 4000 Hz speech bandwidth was then divided into 16 sub-bands, each containing 250 Hz. A threshold was set to delete redundancies. The threshold level in this example was fixed at 10% of the maximum magnitude of each frame. If any of the frequency components in a sub-band was above the threshold, this sub-band was kept otherwise it was removed as redundant. In the first frame, sub-bands 1,2,3,7,8,9,10 and 11 were retained and the rest were eliminated. If one bit is used for each sub-band as control information, and '1' for retaining and '0' for deleting, the control bits for the first frame are 1110001111100000. In second frame, sub-band

1.2.3.8.9.10 and 11 were kept, so the control bits are 1110000111100000. For convenience, only Frequency Companding was used to compress the speech. The compressed speech waveforms are shown in Figure 6.15 and the value of the samples of the compressed speech are shown in Table 6.5. The sampling frequency is 4000 Hz for the first frame and 3500 Hz for the second frame. The samples taken at these sampling rates are shown as crosses on the waveform (Figure 6.15) and their values are those shown in bold type in Table 6.5. The stored data in the Hex format are shown in Table 6.6 with the control bits in bold type. Each byte represents one sample. The control bits occupy the first two bytes in each frame. During the reconstruction, the control bytes for the first frame, *E3* and *E0*, were read into the system. It provided adequate information about this frame, the spectrum structure, the sampling rate and the number of samples stored. Here the sampling rate was 4000 Hz and 64 samples were taken. The system then read in the next 64 bytes and started to reconstruct the speech. For the next frame, the system read in the first two bytes from the remaining data to start the reconstruction. The same process was carried on until all the speech had been reconstructed.

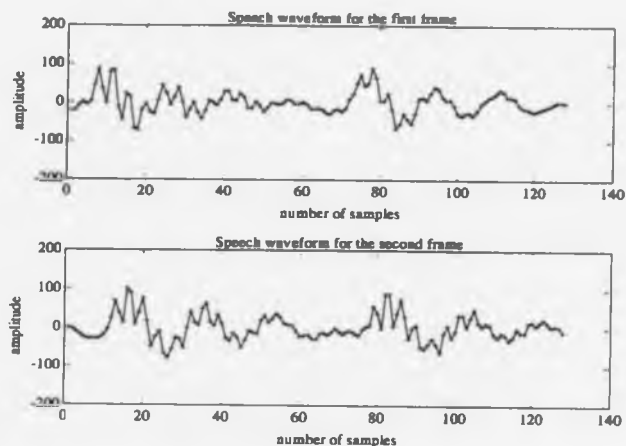


Figure 6.14: Speech waveform in two 16-ms frames.

-21	-21	-7	1	-6	3	46	87	35	-2	80	84
-10	-44	23	15	-66	-72	-18	-5	-24	-29	7	45
25	-6	13	39	3	-39	-17	0	-25	-42	-22	6
0	-7	7	28	30	7	4	24	14	-14	-15	2
-8	-23	-13	-1	-2	-6	-2	7	6	-4	-4	1
-5	-16	-16	-15	-19	-27	-30	-20	-17	-23	-15	9
22	50	73	44	49	90	65	5	4	23	-18	-68
-56	-29	-43	-55	-29	9	13	7	25	41	35	14
6	7	-7	-28	-34	-28	-26	-33	-22	-5	4	12
17	24	33	28	14	14	10	-6	-14	-15	-21	-23
-19	-16	-12	-8	-3	2	2	0				
-1	-3	-10	-18	-25	-29	-28	-30	-28	-20	-6	24
69	39	13	101	90	9	39	76	17	-50	24	-9
-65	-78	-55	-25	-30	-52	0	42	14	7	49	63
16	5	34	11	-26	-34	-15	-24	-51	-32	-7	-12
-15	18	31	15	25	37	28	12	8	7	-8	-22
-20	-18	-28	-29	-15	-13	-19	-11	-1	-9	-16	-9
-8	-14	-20	-5	3	8	56	39	-1	87	89	5
39	74	28	-28	0	8	-48	-54	-42	-27	-42	-64
-13	6	-25	-9	34	34	2	27	48	15	4	12
8	-15	-24	-13	-18	-32	-23	-2	-11	-13	12	17
6	14	23	13	3	6	4	-10				

Table 6.4: The samples of the speech signal for frame 1 and 2

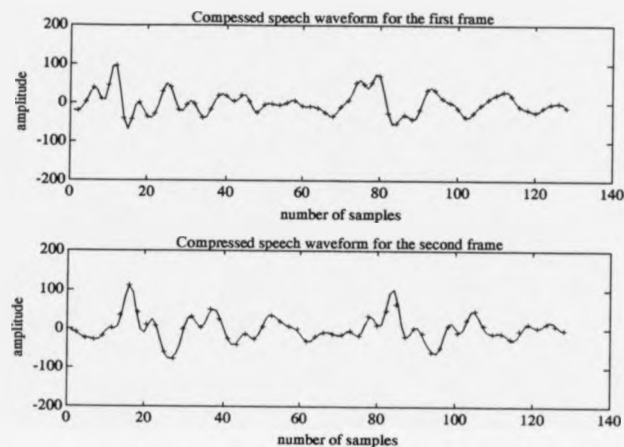


Figure 6.15: Compressed speech waveform in two 16-ms frames.

-20	-21	-13	2	24	39	32	10	8	45	94	96
34	-40	-68	-42	-6	0	-18	-36	-39	-28	-3	29
50	41	9	-18	-21	-5	5	-3	-24	-39	-36	-17
4	19	23	18	9	4	9	20	20	4	-17	-25
-16	-5	-1	-3	-5	-6	-4	0	6	6	-1	-8
-10	-9	-10	-13	-18	-25	-33	-35	-24	-10	-1	7
27	52	60	48	38	52	74	70	27	-26	-55	-53
-38	-31	-35	-42	-39	-18	10	33	40	33	21	11
5	1	-3	-14	-28	-37	-36	-26	-15	-6	1	10
17	21	25	29	29	18	2	-9	-14	-17	-21	-24
-20	-12	-6	-3	-1	0	-2	-12				
-4	-8	-15	-21	-24	-25	-28	-27	-18	-4	2	0
6	35	82	111	95	43	-2	-10	11	25	9	-27
-60	-76	-77	-65	-38	-1	27	31	16	3	11	33
50	47	26	-1	-25	-40	-40	-27	-14	-13	-21	-26
-15	6	27	36	34	27	19	12	7	4	-1	-12
-25	-32	-30	-20	-13	-9	-9	-10	-13	-16	-15	-8
-5	-11	-17	-9	14	33	28	10	12	47	92	101
63	8	-20	-14	2	4	-11	-31	-48	-60	-62	-48
-18	9	16	2	-9	-1	23	44	46	30	9	-5
-11	-12	-10	-11	-16	-25	-29	-21	-5	9	13	8
3	5	13	19	17	8	0	-3				

Table 6.5: The samples of the compressed speech signal for frame 1 and 2.

E3	E0	EB	02	27	0A	2D	60	D8	D6	00	DC
E4	1D	29	EE	FB	FD	D9	EF	13	12	04	14
04	E7	FB	FD	FA	00	06	F8	F7	F3	E7	DD
F6	07	34	30	34	46	E6	C6	E1	D6	EE	21
21	0B	01	F2	DB	E6	FA	0A	15	1D	12	F7
EF	E8	F4	FD	00	F4	E1	E0	F8	E8	E4	EE
02	23	6H	2B	0B	09	C4	B3	FF	1F	03	32
1A	E7	D8	F3	E6	06	22	13	07	FF	E0	EC
F7	F3	F1	FB	EF	21	0A	2F	3F	EC	02	F5
C4	D0	09	F7	17	2F	09	F4	F5	E7	FB	0D
03	0D	08	FD								

Table 6.6: The samples of the compressed speech signal for frame 1 and 2 in Hex format.

## 6.5 Real-Time ARS Using DSP

This section describes an example of real-time speech compression - Frequency Companding - using a Digital Signal Processor. An IBM PC and an on board DSP32C digital signal processor were used to implement the Frequency Companding technique in real time. The speech data processing was carried out in both time domain and frequency domain. The whole process was carried out in several steps. In the first step, the on board A/D converter took the samples from the input speech at the Nyquist rate of 8000 Hz and digitised them with an 8-bit PCM. The second step divided the incoming speech samples into short time frames, with each of the frames having a duration of 16 ms, and containing 128 samples. Each frame was then Fourier transformed to obtain its frequency response. The third step carried out the redundancies elimination and bandwidth compression. The threshold level was pre-set at a certain level according to the requirements and could be

reset at any time without interrupting the process. In fourth step, the compressed version of the speech signal was transformed back to its time domain waveform using inverse FFT. The final step at the transmitter took a sub-set of samples from the spectrally compressed speech samples at a low sampling rate and transmitted them. At the receiving end, the receiver first took FFT of the incoming signal to obtain its frequency domain components. In second step, the compressed bands were restored to their original place. The third step took the inverse FFT to convert the speech signal back into time domain. Finally the speech was converted back to an analogue waveform via a D/A converter, and played back over a loudspeaker.

The DSP32C processing board is a powerful DSP application system. The complete analogue I/O subsystem managed by the DSP32C is dedicated to real time manipulation of sampled data. Operating at 50 MHz, it processes a maximum of 25 million floating point multiply/accumulate instructions per second. The aim of this investigation is to ascertain whether or not the capabilities of the DSP32C are sufficient to handle all the processes involved in implementing the speech frequency companding software, while maintaining a real time operation. The main question to be answered is whether or not the processor, having sampled the speech, can perform an FFT algorithm (to translate the time varying signal into frequency responses), implement the frequency companding algorithm, perform the corresponding inverse FFT (IFFT) (translate the frequency responses back to time domain waveform) and then transmit the result (on to a suitable o/p channel) in a given time.



The input speech was sampled at its Nyquist rate of 8000 Hz. When the incoming samples filled up a buffer which has a size of one analysis frame, here it was 16 ms, or 128 samples, the compression processing started. At the same time new samples were fed into the buffer, so the maximum allowable time for compression of the speech data to take place was the time period of one analysis frame. The major time consuming is that for the FFT and IFFT routines. For the frame size of 16 ms, the execution time for 128 point FFT and IFFT are 0.468 ms each [AT&T WE DSP32C]. This processing time falls far short of the frame length of 16 ms. When the compression processing has finished for one frame of speech samples, the processor is set in a wait state until it receives a buffer full flag from the interrupt service routine (ISR) acknowledging that another full frame (128 samples) has been successfully read in and is now ready for processing. The total time delay of this system is, therefore, one frame period, 16 ms. A timing diagram that shows the sequences of operation of the active software can be seen in Figure 6.16.

The time delay at the receiver side is the same as that in at the transmitter. The total time delay of the transmission system is therefore the duration of two frames. Experimental results show that at a frame size less than 64 ms, the delay is almost unnoticeable. When the frame size increases to more than 128 ms, however, the delay may become significant.

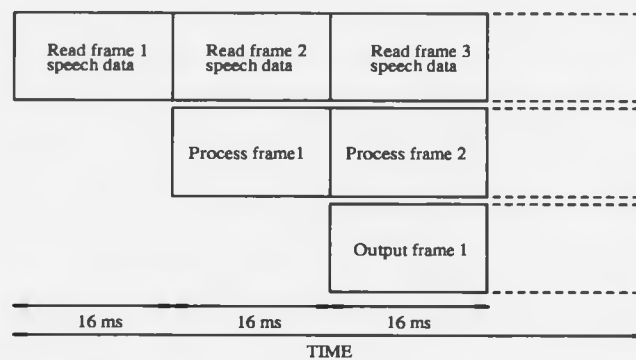


Figure 6.16: DSP timing diagram

## **Chapter 7**

# **Conclusion and Suggestions for Further Work**

## 7.1 Conclusion

The research described in this thesis investigated the use of adaptive-rate sampling in digital speech processing to achieve efficient and reliable speech transmission and storage. A secure speech transmission scheme using the adaptive-rate sampling was also studied. This chapter reviews the techniques developed and suggests areas where further work could be carried out.

### 7.1.1 Adaptive-Rate Sampling for Speech Transmission

The adaptive-rate sampling techniques, namely Time Domain Compression and Frequency Companding, developed in Chapter 3 have wide applications in digital speech transmission and storage. Both of these techniques are based on the premise that the inherent redundancies in the human speech can be removed without significant loss of speech quality. The threshold for the redundancy elimination can be adjusted so that the ratio of the speech compression and the resultant speech quality can be controlled. The ability to control threshold level makes the adaptive-rate sampling technique very useful in low rate speech transmission. In practice, the sampling rate can be fixed as desired and the threshold can then be adjusted to a level at which the compression process will delete a certain amount of the trivial frequency components so that the retained frequency spectrum poses a structure to which the desired sampling rate can be applied.

The combined adaptive-rate sampling and speech coding techniques demonstrated the great potential for a reduction in the transmission rate. Most of the waveform coding techniques and transform coding techniques can be applied to the compressed speech. This has been shown in the examples of combined ARS with Hadamard coding and APCM.

The primary study in combining adaptive-rate sampling with error control coding for a reliable speech transmission in Gaussian channel shows a useful method of extending the usable communication channel. The flexible adaptive system can also be applied to time-variable channels, where the transmission conditions vary rapidly with time, such as mobile communication in which the the channel characteristics depend on the position of the mobile unit.

The adaptive-rate sampling techniques are simple and can be easily implemented in real time using digital signal processing techniques.

### **7.1.2 Low Rate Speech Transmission Using Frame Differences**

The adaptive-rate sampling algorithms developed in Chapter 3 provided the basis of the sampling reduction for multi-band-pass signals. In Chapter 4, a speech signal manipulation method was proposed and investigated. Different from the conventional speech transmission, this technique takes the advantage of pseudo-periodic property in speech signal to achieve a low rate

transmission. The signal to be transmitted is the extracted differential signal between two neighbouring frames. Due to the similarity between two frames of speech, the differential signal will usually occupy a very narrow bandwidth. By using the adaptive-rate sampling techniques, a very low sampling rate can be achieved. The pitch detection plays a vital role in this technique. Precise detection of the pitch in voiced speech provides the basis for the differential signal extraction and, therefore, a low rate sampling. Speech coding techniques can also be employed to the frame differential signal to reduce transmission rate further after the sampling rate reduction.

### **7.1.3 Speech Scrambling Using Adaptive-Rate Sampling**

The speech scrambling technique described in Chapter 5 scrambles the original speech spectrum in the defined patterns to which the adaptive-rate sampling process can be employed to recombine the spectrum allows recovery of the original signal. In this technique, there are two keys for the speech reconstruction, the spectrum structure and the sampling frequency. Firstly the spectrum structure information allows the receiver to remove the unwanted frequency components introduced by the process of low rate sampling. Secondly the correct sampling rate will then recombine the retained spectrum to form the desired speech spectrum which make up an intelligible speech. The changes of the scrambling pattern and sampling rate from frame to frame

gives the system a high degree of security. Multi-user scrambling makes an efficient use of the communication channel. The transmission has to be carried out in a systematic manner to ensure that the same scrambling pattern will not be used by two or more users at the same time and that there is no overlap between any of the scrambled spectrum. This technique gives a very low residual intelligibility in the scrambled speech and can be implemented in real time using digital signal processors.

## **7.2 Suggestions for Further Studies**

The adaptive-rate sampling has wide range of application in low rate transmission, storage and reliable speech transmission etc. Some further work which could be pursued is presented in the next section.

### **7.2.1 Combine Low Rate Sampling With Parameter Coding**

The combined low rate sampling and speech coding techniques show a great potential for a reduction in the transmission rate. The research carried out so far is only that for waveform coding and transform coding. For parameter coding, such a combination is not straight forward. Unlike waveform coding and transform coding which encode the speech samples or transformed format of the speech samples, parameter coding extracts vocal parameters from the input speech and encodes and transmits these parameters. The choice of

parameters is based on statistical study of speech characters, such as VUS pattern, formants, pitch etc. In the adaptive-rate sampling, these characters will be changed after the low rate sampling procedures. It is not possible to apply parameter coding techniques directly to the compressed version of speech. However, a modified parameter coding technique could be applied to the compressed speech. In spite of the changes of characters, the compressed speech signal is still a speech related signal; and has strong relation with the original speech. A further study can be carried out in order to work out the statistical relation between the compressed speech signal and the parameters needed for the coding techniques.

### **7.2.2 Speech Compression Using Differential Signal in Time Domain**

In the technique of reconstructing speech from frame differences, the differential signal was extracted by making a comparison between two consecutive frame of speech spectrum. A similar process could be applied to time domain. When a voiced speech is detected, the length of the pitch period  $P_l$  and the number of repetition of the pitch period  $P_r$  can be counted. The detection of the repetition should be carried out by making comparison between the first frame signal and each following frame until the mismatch exceeds a defined matching threshold. Such comparison can be conducted in both time domain and frequency domain. Due to the pseudo periodic property, two



neighbouring period signals should have a similar waveform structure with possible amplitude variation. The average amplitude variation  $M_a$  usually indicates the change of speech energy. The information to be encoded and transmitted consists of  $P_l$ ,  $P_r$  and  $M_a$  for each of the  $P_r$  frames and the speech samples in the first period. At the receiver, the first period speech is repeated  $P_r$  times and every time it is modified by a factor of  $M_a$ . In this method, the matching threshold and the amplitude variation should play the key role for the speech reconstruction. A generalised, effective matching threshold should be defined. An algorithm should be developed to calculate the average amplitude variation accurately.

### 7.2.3 Reliable Transmission of Speech

The primary study described in Chapter 6 shows some encouraging progress in reliable speech transmission. Although the simulation results were obtained on a Gaussian channel, an improvement of transmission performance is expected in other types of channel which suffer from different distortion effects, such as fading, interference etc. One practical channel, in which the adaptive error control technique is a powerful tool to establish a reliable transmission, is the high frequency (HF) channel. The HF channel characterised by fading, interference and multipath is an extreme example of a time variable channel and is unsuitable for reliable long term transmission without some kind of checking and adaptation of the transmission method at regular intervals. The RTCE technique has been used to detect

the channel condition in HF communications and can be used to control the adaptation of the error control coding scheme and the sampling rate. Many appropriate RTCE procedures have been developed [Darnell.(1983)] [Honary, *et al.*(1988)] [Zolghadr, *et al.*(1988.a)] [Zolghadr, *et al.*(1988.b)] and should be evaluated comparatively in the further study.

Another suggestion for the further work on reliable transmission is that using Unequal Error Protection (UEP) codes [Boyarinov, *et al.*(1981)]. The speech transmission is basically a numerical data transmission in which errors in the sign or in the high-order digits are more serious than the errors in the low-order digits. The unequal important level in transmitted bits stream is more obvious in the adaptive rate sampling techniques. Here, the control information added for speech reconstruction is extremely important, an error in the control information could cause complete failure of the reconstruction. By using UEP, these information can have a high degree of protection. For the speech information, high-order digits should be protected more than the low-order digits.

#### **7.2.4 Noise Reduction Using Dynamic Thresholding**

The dynamic threshold process described in Chapter 3 was initially used to remove inherent redundancies from a speech signal. It can also provide a means of noise reduction from speech. In a Gaussian channel, the transmitted speech is corrupted by the added noise. This noise has a flat spectrum and usually this spectrum is much lower than of speech signal. By putting a

threshold at the received speech spectrum, some of the noise frequency components can be removed. Obviously, the threshold level has to be adjusted according to the channel condition. This method is only suitable in certain channel conditions. When the channel noise is low, a very low threshold level is enough to remove the noise. But when the channel noise become significant, the thresholding process which is there to delete noise will also delete significant amount of speech information, which gives a dramatic degradation of the reconstructed speech quality. A further study should be conducted to investigate the relationship between the speech signal, channel condition and the threshold level.

## Bibliography

- [Reeves,(1938)] Reeves, A. H.: French Patent 852183, 1938.
- [Shannon,(1949)] Shannon, C.E.: "Communications in the Presence of Noise", *Proc. IRE*, vol.37, pp.10-21, January 1949.
- [Cutler,(1952)] Cutler, C.: "Differential Quantization of Communications", U.S. Patent 2,605,361, July 1952.
- [DeJager,(1952)] DeJager, F.: "Delta Modulation, a Method of PCM Transmission Using a 1-unit Code", *Philips Res. Rep.*, pp.442-466, December 1952.
- [Elias,(1955)] Elias, P.: "Coding for Noise Channels", *IRE Nat. Conv Record*, vol.3, pt.4, pp.37-46, 1955.
- [Smith,(1957)] Smith, B.: "Instantaneous Companding of Quantized Signals", *The Bell Syst. Tech. J.*, pp.653-709, 1957.
- [Linden,(1959)] Linden, D.: "A Discussion of Sampling Theorems" *Proc. of the IRE*, pp.1219-1226, July 1959.

- [Max,(1960)] Max, J.: "Quantizing for Minimum Distortion", *IRE Trans. on Inform. Theory*, vol.IT-6, pp.7-12, March 1960.
- [Bogert,(1963)] Bogert, B., Healy, M. and Tukey, J.: "The Quefrency Analysis of Time Series for Echoes", *Proc. Symp. on Time Series Analysis*, M. Rosenblatt, Ed., Ch.15, pp.209-243, J. Wiley, New York, 1963.
- [Bogner,(1965)] Bogner, R.E.: "Frequency Division in Speech Bandwidth Reduction", *IEEE Trans. on Comm.*, vol.COM-13, no.4, pp.438-451, December 1965.
- [Abate,(1967)] Abate, J.: "Linear and Adaptive Delta Modulation", *Proc. IEEE*, vol.55, pp.298-308, March 1967.
- [Kahn,(1967)] Kahn, D.: "The Code-Breakers", New York: Macmillan, 1967.
- [Boesswetter,(1970)] Boesswetter, C.: "Analog Sequence Analysis and Synthesis of Voice Signals", *Proc. Symp. on Applied Walsh Functions*, Washington D.C. pp.220-229, 1970.
- [Kelly,(1970)] Kelly, L.C.: "Speech and Vocoder", *Radio and Electronic Engineer*, vol.40(2), pp.73-82, August 1970.
- [Robinson,et al.(1970)] Robinson, G.S. and Granger, R.L.: "Fast Fourier Transform Speech Compression", *Proc. 1970 IEEE Int. Conf. Comm.*, paper 26-5, June 1970.

- [Schafe,*et al.*(1970)] Schafe, R.W. and Rabiner, L.R.: "System for Automatic Formant Analysis of Voiced Speech", *J. Acoust. Soc. Am.*, vol.47, pp.634-648, February 1970.
- [Atal, *et al.*(1971)] Atal, B. and Hanauer, S.: "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Am.*, vol.50, pp.637-655, 1971.
- [Campanella,*et al.*(1971.a)] Campanella, S.J. and Robinson, G.S.: "A Comparison of Walsh and Fourier Transformations for Application to Speech", *Proc. 1971 Symp. Walsh Functions*, Washington, D.C. pp.199-205, 1971.
- [Campanella,*et al.*(1971.b)] Campanella, S.J. and Robinson, G.S.: "A Comparison of Orthogonal Transformation for Digital Speech Processing", *IEEE Trans. on Comm.*, vol.COM-19, pp.1045-1049, 1971.
- [Robinson,(1971)] Robinson, G.S.: "Walsh-Hadamard Transform Speech Compression", *Proc. 4th Hawaii Int. Conf. System Sciences*, pp.411-413, June 1971.
- [Rosenberg,*et al.*(1971)] Rosenberg, A., Schafer, R. and Rabiner, L.: "Effects of Smoothing and Quantizing the Parameters of Formant-Coded Voiced Speech", *J. Acoust. Soc. Am.*, vol.50 no.6, pp.1532-1538, December 1971.

- [Viterbi,(1971)] Viterbi, A.: "Convolutional Codes and Their Performance in Communication Systems", *IEEE Trans. on Comm.*, vol.COM-19, pp. 751-772, October 1971.
- [Blessner,(1972)] Blessner, B.: "Speech Perception Under Conditions of Spectral Transformation: I. Phonetic Characteristics", *J. Speech and Hearing*, vol.15, pp.5-41, 1972.
- [Flanagan,(1972)] Flanagan, J.L.: "Speech Analysis, Synthesis and Perception". New York, Springer-Verlag, 1972.
- [Markel,(1972)] Markel, J.D.: "The SIFT Algorithm for Fundamental Frequency Estimation", *IEEE Trans. on Audio Electroacoust.*, vol.AU-20, pp.367-377, December 1972.
- [Pacz *et al.*(1972)] Pacz, M. and Glisson, T.: "Minimum Mean Squared-Error Quantization in Speech, PCM and DPCM Systems", *IEEE Trans. on Comm.*, vol.COM-20, pp.225-230, April 1972.
- [Cummiskey,*et al.*(1973)] Cummiskey, P., Jayant, N. and Flanagan, J.: "Adaptive Quantization in Differential PCM Coding of Speech", *The Bell Syst. Tech. J.*, pp.1105-1118, September 1973.
- [French.(1973)] French, R.: "Speech Scrambling and Synchronization", *Philips Res. Rep.*, no.9, pp.1-115, 1973.

- [Levitt,(1973)] Levitt, H.: "Speech processing aids for the deaf: An overview", *IEEE Trans. on Audio and Electroacoustics*, vol.AU-21, pp.269-273, June 1973.
- [Markel, et al.(1973)] Markel, J. and Gray Jr., A.: "On Autocorrelation Equations as Applied to Speech Analysis", *IEEE Trans. on Audio and Electroacoustics*, vol.AU-21, pp.69-79, April 1973.
- [Shum,et al.(1973)] Shum, Y.Y., Elliot, A.R. and Brown, O.W.: "Speech Processing with Walsh-Hadamard Transform", *IEEE Trans. on Audio and Electroacoustics*, vol.AU-21, pp.174-179, June 1973.
- [Wilkinson.(1973)] Wilkinson, R.M.: "A 4800 bits/s Adaptive PCM Speech Coder", *SRDE Report*, no.7303, September 1973.
- [Jayant,(1974)] Jayant, N.S.: "Digital Coding of Speech Waveform: PCM, DPCM and DM Quantizers", *Proc. of IEEE*, pp.611-632, May 1974.
- [McCalmont,(1974)] McCalmont, A.M.: "How to select and apply various voice scrambling techniques", *Communications News*, pp.34-37, January 1974.
- [Ross,et al.(1974)] Ross, M.J., Shaffer, H.L., Cohen, A., Freudberg, R and Manely, H.J.: "Average Magnitude Difference Function Pitch Extractor", *IEEE Trans. on Acoust. Speech Signal Processing*, vol.ASSP-22, pp.353-362, October 1974.



- [Makhoul,(1975)] Makhoul, J.: "Linear Prediction: a tutorial review", *Proc. of the IEEE*, vol.63, no.4, pp.561-580, April 1975.
- [Miller,(1975)] Miller, N.J.: "Pitch Detection by Data Reduction", *IEEE Trans. on Acoust. Speech, Signal Processing*, vol.ASSP-23, pp.72-79, February 1975.
- [Rosenber,*et al.*(1975)] Rosenber, A.E. and Sambur, M.R.: "New Techniques for Automatic Speaker Verification", *IEEE Trans. on Acoust. Speech, Signal Processing*, vol.ASSP-23, pp.169-176, April 1975.
- [Atal, *et al.*(1976)] Atal, S.B. and Rabiner, L.R.: "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition", *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-24, no.3, pp.201-212, June 1976.
- [Dubnowsk,*et al.*(1976)] Dubnowsk, J.J., Schafer, R.W. and Rabiner, L.R.: "Real-Time Digital Hardware Pitch Detector", *IEEE Trans. on Acoust, Speech Signal Processing*, vol.ASSP-24, pp.2-8, 1976.
- [Flanagan,(1976)] Flanagan, J.L.: "Computers That Talk and Listen: Man-Machine Communication by Voice", *Proc. IEEE*, vol.64, no.4, pp.416-432, April 1976.
- [Rabiner, *et al.*(1976)] Rabiner, L.R, Cheng, M.J, Rosenberg, A.E and McGonegal, C.A.: "A Comparative Performance Study of Several Pitch

- Detection Algorithms", *IEEE Trans. on Acoust. Speech. and Signal Processing*, vol.ASSP-24, no.5, October 1976.
- [Nelson,(1976)] Nelson, R.: "Guide to Voice Scramblers for Law Enforcement Agencies", National Bureau of Standards, Washington, D.C. 20234, December 1976.
- [Turner,(1976)] Turner, L.F.: "Application of Data Compression to an Experimental 9.6 kb/s Adaptive PCM Digital Speech System", *Proc. IEE*, vol.123(2), pp.109-114, February 1976.
- [Baschlin,(1977)] Baschlin, W.: "The Integration of Time Division Speech Scrambling into Police Telecommunication Networks", *Proc. Carnhan Conf. on Crime Countermeasures*, pp.141-145, 1977.
- [Crochiere,(1977)] Crochiere, R.: "On the Design of Sub-Band Coders for Low-Bit-Rate Speech Communication", *The Bell Syst. Tech. J.*, vol.56, pp.747-770, May-June 1977.
- [Frangoulis,(1977)] Frangoulis, E. and Turner, L.F.: "Hadamard-transformation Technique of Speech Coding: Some Further Results", *Proc. IEE*, vol.124(10), pp.845-852, October 1977.
- [Gold,(1977)] Gold, B.: "Digital Speech Networks", *Proc. of IEEE*, vol.65, no.12, pp.1638, December 1977.

- [Kak, et al.(1977)] Kak, S.C. and Jayant, N.S.: "On Speech Encryption Using Waveform Scrambling", *The Bell Syst. Tech. J.*, vol.56, pp.781-808, May-June 1977.
- [Leitich,(1977)] Leitich, A: "Scrambler Design Criteria", *Proc. Carnhan Conf. on Crime Countermeasures*, pp.5-9, 1977.
- [Belland,et al.(1978)] Belland, E. and Bryg, N.: "Speech Signal Privacy System Based on Time Manipulation", *Proc. Carnahan Conference on Crime Countermeasures*, Univ. of Kentucky, 1978.
- [Braustad,(1978)] Braustad, D.: "Security of Computer Communication", *IEEE Comm. Soc. Mag.*, November 1978.
- [Orceyre,et al.(1978)] Orceyre, M. and Heller, R.: "An Approach to Secure Voice Communication Based on the Data Encryption Standard" *IEEE Comm. Soc. Mag.*, November 1978.
- [Rabiner,et al.(1978)] Rabiner, L.R. and Schafer, R.W.: "Digital Processing of Speech Signal", *Prentice-Hall*, 1978.
- [Alexander,(1979)] Alexander, J.D.H.: "Speech Goes Digital", *Racal Review*, pp.5, Autumn, 1979.
- [Cain,et al.(1979)] Cain, J., Clark, G. and Geist, J.: "Punctured Convolution Codes of Rate  $(n-1)/n$  and Simplified Maximum Likelihood Decoding", *IEEE Trans. on Inf. Theory*, vol.IT-25, pp.97-100, January 1979.

- [Brunner,(1980)] Brunner, E.R.: "Efficient scrambling techniques for speech signals", *Proc. Int. Conf. Comm.*, Seattle, WA, pp.16.1.1-16.1.6, June 1980.
- [Gieseler,et al.(1980)] Gieseler, P.B. and O'Neal, J.B. Jr.: "Speech Bandwidth Reduction", *FCC Working Paper*, November 1980.
- [Boyarinov, et al.(1981)] Boyarinov, I. and Katsman, G.: "Linear Unequal Error Protection Codes", *IEEE Trans. on Inf. Theory*, vol.IT-27, no.2, pp.168-175, March 1981.
- [Hong, et al.(1981)] Hong, S. and Kuebler, W.: "An Analysis of Time Segment Permutation Methods in Analog Voice Privacy Systems", *Proc. Carnahan Conf. on Crime Countermeasures*, University of Kentucky, 1981.
- [Jayant, et al.(1981)] Jayant, N.S., McDermott, B.J., Christensen, S.W and Quinn, A.M.: "A Comparison of Four Methods for Analog Speech Privacy", *IEEE Trans.on Comm.*, vol.COM-29, pp.18-23, January 1981.
- [Malah,et al.(1981)] Malah, D. and Flanagan, J.L.: "Frequency Scaling of Speech Signal by Transform Technique", *The Bell Syst. Tech. J.*, vol.60, no.9, pp.2107-2150, November 1981.
- [Crochiere,et al.(1982)] Crochiere, R.E., Cox, R.V. and Johnston, D.: "Real-Time Speech Coding", *IEEE Trans. on Comm.*, vol.COM-30, pp.621-634, April 1982.

- [Jayant.(1982)] Jayant, N.S.: "Analog Scramblers for Speech Privacy", *Computer Security*, vol.1, pp.275-289, 1982.
- [Kitawaki, *et al.*(1982)] Kitawaki, N., Itoh, K., Honda, M and Kakehi, K.: "Comparison of Objective Speech Quality Measures for Voice-band Coders", *Proc. Int. Conf. Acoust. Speech Signal Processing*, vol.2, pp.1000-1003, 1982.
- [Ramstad,(1982)] Ramstad, T.A.: "Sub-band Coder with a Simple Adaptive Bit Allocation Algorithm", *Proc. Int. Conf. Acoust. Speech Signal Processing*, pp.203-207, 1982.
- [Sambur,(1982)] Sambur, M.R.: "Speech Algorithm Advances Promise Toll-quality Medium-Band Digitised Speech", *Speech Technology*, vol.1(3), pp.22-34, September/October 1982.
- [Wong,*et al.*(1982)] Wong, W.C., Steele, R. and Xydeas, C.S.: "Transmitting Data on the Phase of Speech Signals", *The Bell Syst. Tech. J.*, vol.61, no.10, pp.2947-2970, December 1982.
- [Darnell.(1983)] Darnell, M.: "Real-Time Channel Evaluation", *AGARD Lecture Series*, no.127 on Modern HF Communications, 1983.
- [Goodman, *et al.*(1983)] Goodman, D.J. and Sundberg, C.E.: "Combined Source and Channel Coding for Variable-Bit-Rate Speech Transmission" *The Bell Syst. Tech. J.*, September 1983.

- [Jayant, *et al.*(1983)] Jayant, N.S., Cox, R.V., Mcdermott, B.J. and Quinn, A.M.: "Analog Scrambling for Speech Based on Sequential Permutations in Time and Frequency", *The Bell Syst. Tech. J.*, January 1983.
- [Patrick, *et al.*(1983)] Patrick, P.J., Steele, R. and Xydeas, C.S.: "Frequency Compression of 7.6 kHz Speech into 3.3 kHz Bandwidth", *IEEE Trans. on Comm.*, vol.COM-31, no.5, pp.692-701, May 1983.
- [Lee, *et al.*(1984)] Lee, L., Chou, G and Chang, C.: "A New Frequency Domain Speech Scrambling System Which Does Not Require Frame Synchronization", *IEEE Trans. on Comm.*, vol.COM-32, no.4, pp.444-456, April 1984.
- [Dodson, *et al.*(1985)] Dodson, M.M. and Silva, A.M.: "Fourier Analysis and the Sampling Theorem", *Proc. Royal Irish Acad.* vol.85A, pp.81-108, 1985.
- [Honda, *et al.*(1985)] Honda, M. and Itakura, F.: "Bit Allocation in Time and Frequency Domain for Predictive Coding of Speech", *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-32, pp.465-473, June 1985.
- [Jibbe,(1986)] Jibbe, M.K.: "Compression System for Minimizing Space Requirement for Storage and Transmission of Digital Speech Signal", *Ph.D Dissertation*, Wichita University, 1986.

- [Soong, *et al.*(1986)] Soong, F.K., Cox, R.V. and Jayant, N.S.: "A High Quality Subband Speech Coder with Backward Adaptive Predictor and Optimal Time-Frequency Bit Assignment", *Proc. Int. Conf. Acoust. Speech Signal Processing*, pp.2387-2390, 1986.
- [Parsons,(1987)] Parsons, T.: "Voice and Speech Processing", *McGraw-Hill*, 1987.
- [Dodson, *et al.*(1988)] Dodson, M.M. and Silva, A.M.: "An Algorithm for Optimal Sampling", *Signal Processing*, 1988.
- [Hole.(1988)] Hole, K.: "New Short Constraint Length Rate  $(N-1)/N$  Punctured Convolutional Codes for Soft-Decision Viterbi Decoding", *IEEE Trans. on Inf. Theory*, vol.IT-34, no.5, September 1988.
- [Holmes,(1988)] Holmes, J.N.: "Speech Synthesis and Recognition", *Van Nostrand Reinhold Co. Ltd.* 1988.
- [Honary, *et al.*(1988)] Honary, B., Shaw, M. and Darnell, M.: "A New ARQ Transmission Scheme Involving Zero-Crossing Channel Evaluation", *Electronics Letters*, no.10, May 1988.
- [Kitawaki.(1988)] Kitawaki, N.: "Quality Assessment of Speech Coding and Speech Synthesis System", *IEEE Comm. Mag.*, October 1988.
- [Sklar.(1988)] Sklar, B.: "Digital Communication, Fundamental and Applications", *Prentice Hall International Inc.* 1988.

- [Zolghadr, *et al.*(1988.a)] Zolghadr, F., Honary, B. and Darnell, M.: "Statistical Real Time Channel Evaluation (SRTCE) Technique Using Variable Length T-Codes", *Proc. of IEE, Part F, Communications Radar and Signal Processing*, 1988.
- [Zolghadr, *et al.*(1988.b)] Zolghadr, F., Honary, B. and Darnell, M.: "Embedded Convolutional Coding" *Proc. of IERE Conf. on Digital Processing of Signals in Commun.*, September 1988.
- [Del Re, *et al.*(1989)] Del Re, E., Fantacci, R. and Maffucci, D.: "A New Speech Signal Scrambling Method for Secure Communications: Theory, Implementation, and Security Evaluation", *IEEE J. on Selected Areas in Comm.*, vol.7, no.4, May 1989.
- [Jayant, *et al.*(1990)] Jayant, N., Lawrence, V. and Prezaz, D.: "Coding of Speech and Wideband Audio". *AT&T Tech. J.*, September/October 1990.
- [Darnell, *et al.*(1991)] Darnell, M., Honary, B. and He, W.: "Speech Scrambling Employing Adaptive Rate Sampling", *Electronics Letters*, vol.27, no.12, June 1991.
- [Spanias, *et al.*(1992)] Spanias, S. and Wu, F.: "Speech Coding and Recognition: A Review", *IEICE Trans. on Fundamentals of Electronics Commun. & Computer*, vol.E75-A, no.2, February 1992.



---

[AT&T WE DSP32C] "AT&T WE DSP32C Digital Signal Processor Information Manual", *The A&T Documentation Management Organisation*.

[Darnell] Darnell, M.: "Speech Digitisation Techniques".

## List of Symbols and Abbreviation

$\rho$	Optimal Sampling Rate
A/D	Analog to Digital Conversion
ADM	Adaptive Delta Modulation
ADPCM	Adaptive Differential Pulse Code Modulation
APC	Adaptive Predictive Coding
ARS	Adaptive-Rate Sampling
AWGN	Additive White Gaussian Noise
CDMA	Code Division Multiple Access
D/A	Digital to Analog Conversion
DM	Delta Modulation
DPCM	Differential Pulse Code Modulation
DSP	Digital Signal Processor
FC	Frequency Companding
FDMA	Frequency Division Multiple Access
FFT	Fast Fourier Transform
$\rho_m$	Reconstruction Sampling Rate
$\rho_s$	Sampling Frequency
<i>Had</i>	Hadamard Function

HF	High Frequency (2-30 MHz)
Hz	Hertz, frequency in cycles per second
IFFT	Inverse Fast Fourier Transform
LPC	Linear Predictive Coding
MAC	Multiple Access Channel
ms	millisecond
PCM	Pulse Code Modulation
rms	root mean square
RTCE	Real-Time Channel Evaluation
SBC	Sub-Band Coding
SD	Spectrum Distortion
TC	Transform Coding
TDC	Time Domain Compression
TDMA	Time Division Multiple Access
TFSP	Time-Frequency Segment Permutation
TSP	Time Segment Permutation
SIFT	Simplified Inverse Filtering Technique
SNR	Signal to Noise Ratio
$SNR_{seg}$	Segmental Signal to Noise Ratio
VUS	Voiced,Unvoiced,Silence Classification

## Appendix

## List of sets of 10 English language test phrases

### Set 1

Thieves who rob friends deserve jail.  
Open the door but don't break the glass.  
Cats and dogs all hate each other.  
Tread on the grape and make good wine.  
It's always better to be reasonable.  
Never forget that you are British.  
Turn left, then right, then right again.  
Plasma displays are now being used.  
I prefer tea to coffee.  
Paste pink paper on the wall.

### Set 2

All over Holland, the dykes are straight.  
Linear predictive vocoders are best.  
Call back after fifty minutes.  
Stop talking when I am in charge.  
John Brown's body lies a-mouldering in the grave.  
Curtains swish from side to side.  
Dial nine-five-eight in an emergency.  
On multi-stage, finite impulse response filters.  
Colonel Schmidt and Mr. Smith are coming.  
Eight Megahertz should be avoided like the plague.

### Set 3

Name, rank and number please.  
The best months for swimming are June and July.  
E D C using convolutional codes.  
Pencil sharpeners tend to become blunt.  
There was a young lady from Devizes.  
Every packet contains a government health warning.  
Date of birth: two-six-fiftynine.  
Waves on the sea make me sick.  
Feed a cold and starve a fever.  
War insurance is very expensive.

#### Set 4

HF systems are for users to use, not for operators to operate.  
Keep to the centre of the road.  
Twenty-watt lamps give a dim light.  
The Arctic ocean is frozen over.  
The clock on the wall is fifteen minutes fast.  
This test has been carefully devised.  
Change frequency to sixteen point four nine seven.  
Pay the balance or face the consequences.  
Car ferry services operate all the year round.  
Is the earth really flat?

#### Set 5

Such an operation is always efficient.  
The most accurate measurement is that of frequency.  
Would you put your name and address.  
You make me feel as if I forgot something.  
It has been common practice to share.  
He is a member of many organisations.  
For example, take 17 from 30.  
She is a respectable business woman.  
There are large amounts of natural noise.  
We are concerned at the lack of support.

#### Set 6

Look for the sign to the Concert Hall.  
If you knew more, you would understand better.  
The Committee shall have only 5 members.  
Take care when you cross the road.  
Assessing results is a difficult process.  
The decision she made was an obvious one.  
Values will be expressed as percentages.  
America is a very large country.  
You will wish to see how your taxes are spent.

#### Set 7

Give at least 14 days' notice.  
Only one comment at a time please.  
From now on, things will change completely.  
Tell your husband you took me to the garage.  
Problems of special interest in communications.

Key words are given on pages 10 and 50.  
Term ends on the 18th of December.  
It's strange that he did not hear about it.  
Each chapter is followed by several problems.  
Don't write on the wall in pencil.

Set 8

Go past the door marked "Private".  
I think you could answer the question.  
Forward movement can be seen through the glass.  
The sun shines brightly every day.  
People are killing each other without cause.  
This has the advantage of stopping accidents.  
Next Friday is a National holiday.  
Let me see if I can remember the song.  
He was working late at the office.  
You must be wanting a rest from all this.



## List of sets of 40 English language test words

Set 1	Set 2	Set 3	Set 4	Set 5
wish	xylophone	yesterday	yet	abbreviate
abroad	almost	analyse	answer	party
improvement	income	increase	index	hotel
passenger	penalty	pitch	plan	rapid
balance	bank	base	blame	double
standard	state	stock	strong	garden
time	tourist	trade	training	gravity
quality	question	rapid	relative	sit
money	month	moon	name	control
five	following	food	forecast	lack
correct	count	culture	decide	effect
declare	defect	demonstrate	describe	zero
plant	pleasant	position	possible	kidnap
visit	walk	weight	wild	fathom
anxious	arrange	assembly	average	seem
gentle	glass	good	government	market
exact	example	expand	fair	brandy
simple	since	slight	speak	room
target	telegraph	telephone	think	forecast
boat	book	butter	cable	union
study	suppose	survey	system	appeal
dream	election	entertain	equal	improvement
labour	leader	light	little	navy
report	represent	save	second	else
cage	captain	catch	channel	obtain
fashion	fault	finish	firm	counter
treasure	treatment	understand	value	simple
sense	several	shrink	similar	hospital
navy	zoo	obtain	office	liquor
compliment	concern	condition	constant	decide
cheap	city	claim	clear	technical
forward	gain	general	generate	eliminate
oil	outcome	page	party	ton
develop	difference	direct	disclose	identify
speed	stage	stamp	stand	balance
consumer	contact	contract	control	ice
industry	interest	internal	initial	pick
price	problem	producer	protest	capital
market	match	minister	momentum	Japan
previous	ground	hard	head	add